

---

# Lifelong Learning in the Clinical Open World

---

at the Computer Science Faculty  
of the Technischen Universität Darmstadt

approved in fulfillment of the requirements for the degree of  
Doktor-Ingenieur (Dr.-Ing.)

**Doctoral thesis by Camila González**

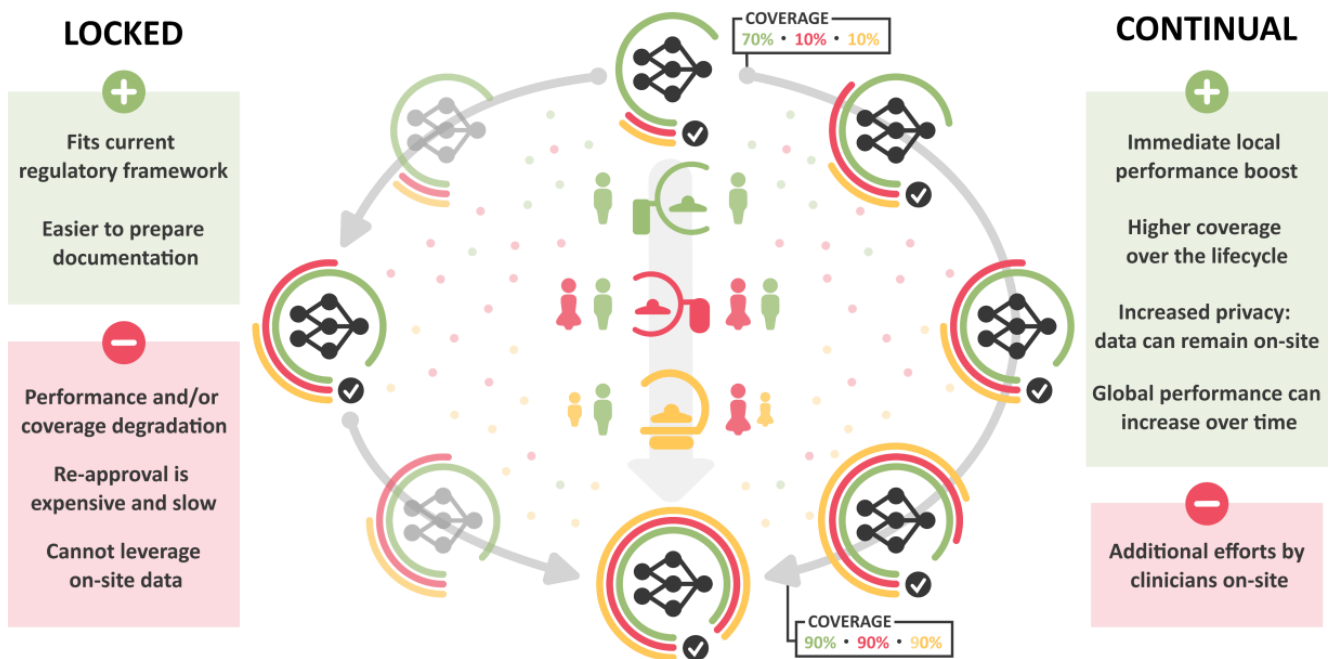
**First assessor:** Anirban Mukhopadhyay, Ph.D.

**Second assessor:** Prof. Dr. techn. Dr.-Ing. eh. Dieter W. Fellner

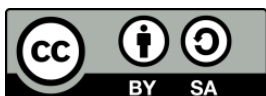
**Third assessor:** Prof. Tianming Liu, Ph.D.

**Darmstadt, 2023**

---



Gonzalez, Camila: Lifelong Learning in the Clinical Open World  
Darmstadt, Technische Universität Darmstadt,  
Year thesis published in TUpriints 2023  
Date of the viva voce 17.03.2023



Published under CC BY-SA 4.0 International  
<https://creativecommons.org/licenses/>

---

## Erklärungen laut Promotionsordnung

---

### **§ 8 Abs. 1 lit. c PromO**

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

### **§ 8 Abs. 1 lit. d PromO**

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

### **§ 9 Abs. 1 PromO**

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

### **§ 9 Abs. 2 PromO**

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 3. Februar 2023

---

C. González

---

# Abstract

---

Despite mounting evidence that data drift causes deep learning models to deteriorate over time, the majority of medical imaging research is developed for – and evaluated on – static close-world environments. There have been exciting advances in the automatic detection and segmentation of diagnostically-relevant findings. Yet the few studies that attempt to validate their performance in actual clinics are met with disappointing results and little utility as perceived by healthcare professionals. This is largely due to the many factors that introduce shifts in medical image data distribution, from changes in the acquisition practices to naturally occurring variations in the patient population and disease manifestation. If we truly wish to leverage deep learning technologies to alleviate the workload of clinicians and drive forward the democratization of health care, we must move away from close-world assumptions and start designing systems for the *dynamic open world*.

This entails, first, the establishment of reliable quality assurance mechanisms with methods from the fields of uncertainty estimation, out-of-distribution detection, and domain-aware prediction appraisal. Part I of the thesis summarizes my contributions to this area. I first propose two approaches that identify outliers by monitoring a self-supervised objective or by quantifying the distance to training samples in a low-dimensional latent space. I then explore how to maximize the diversity among members of a deep ensemble for improved calibration and robustness; and present a lightweight method to detect low-quality lung lesion segmentation masks using domain knowledge.

Of course, detecting failures is only the first step. We ideally want to train models that are reliable in the open world for a large portion of the data. Out-of-distribution generalization and domain adaptation may increase robustness, but only to a certain extent. As time goes on, models can only maintain acceptable performance if they *continue learning* with newly acquired cases that reflect changes in the data distribution. The goal of continual learning is to adapt to changes in the environment without forgetting previous knowledge. One practical strategy to approach this is *expansion*, whereby multiple parametrizations of the model are trained and the most appropriate one is selected during inference. In the second part of the thesis, I present two expansion-based methods that do not rely on information regarding *when* or *how* the data distribution changes.

Even when appropriate mechanisms are in place to fail safely and accumulate knowledge over time, this will only translate to clinical usage insofar as the regulatory framework allows it. Current regulations in the USA and European Union only authorize *locked* systems that do not learn post-deployment. Fortunately, regulatory bodies are noting the need for a modern *lifecycle regulatory approach*. I review these efforts, along with other practical aspects of developing systems that learn through their lifecycle, in the third part of the thesis.

We are finally at a stage where healthcare professionals and regulators are embracing deep learning. The number of commercially available diagnostic radiology systems is also quickly rising. This opens up our chance – and responsibility – to show that these systems can be safe and effective throughout their lifespan.

---

# Zusammenfassung

---

Trotz zunehmender Beweise dafür, dass *Deep Learning* Modelle im Laufe der Zeit an Qualität verlieren, wird der Großteil der Forschung im Bereich der medizinischen Bildgebung in statischen Umgebungen entworfen und evaluiert. Es gab in den letzten Jahren spannende Entwicklungen bei der automatischen Erkennung und Segmentierung diagnostisch relevanter Befunde. Allerdings haben die wenigen prospektiven Studien, die es dazu gab, enttäuschende Ergebnisse gezeigt. Dies ist vor allem auf die vielen Faktoren zurückzuführen, die zu Verschiebungen in der Verteilung medizinischer Bilddaten führen. Diese reichen von Änderungen in den Bildgebungsverfahren bis hin zu natürlich vorkommenden Variationen in der Patientenpopulation und der Ausprägung von Krankheiten. Wenn wir Deep Learning wirksam einsetzen wollen, müssen wir uns von den Annahmen einer geschlossenen Umgebung lösen und damit beginnen, Systeme für die *dynamische offene Welt* zu entwerfen.

Dies erfordert zunächst die Einrichtung zuverlässiger Qualitätssicherungsmaßnahmen. Der erste Teil dieser Dissertation fasst meine Beiträge zu diesem Themengebiet zusammen. Ich schlage zuerst zwei Ansätze vor, welche Ausreißer durch ein selbst überwacht Lernziel oder durch die Quantifizierung des Abstands zu Trainingsbeispielen in einem niedrig dimensionalen latenten Raum identifizieren. Anschließend untersuche ich, wie die Vielfalt unter den Mitgliedern eines tiefen Ensembles maximiert werden kann, um die Kalibrierung und Robustheit zu verbessern. Zudem stelle ich eine domänenbasierte Methode zur Erkennung schlechter Segmentierungsmasken für Lungenläsionen vor.

Natürlich ist die Erkennung von Fehlern nur der erste Schritt. Im Idealfall wollen wir Modelle trainieren, die in der offenen Welt für einen großen Teil der Daten zuverlässig funktionieren. Bisherige Verfahren, unter anderem aus der Domänenanpassung, können zwar die Robustheit erhöhen, aber nur bis zu einem gewissen Grad. Mit der Zeit behalten Modelle nur dann eine akzeptable Leistung bei, wenn sie mit neu erfassten Beispielen weiterlernen, welche die Änderungen in der Verteilung der Daten widerspiegeln. Das Ziel des *kontinuierlichen Lernens* besteht darin, sich an Veränderungen in der Umgebung anzupassen, ohne bereits Gelerntes zu vergessen. Eine praktische Strategie, um dies zu erreichen, ist die *Expansion*, bei der mehrere Parametrisierungen des Modells trainiert werden, und während der Inferenz die am besten geeignete ausgewählt wird. Im zweiten Teil der Arbeit stelle ich zwei Methoden vor, welche auf Expansion basieren, aber nicht auf Informationen darüber angewiesen sind, *wann* oder *wie* sich die Datenverteilung ändert.

Selbst wenn geeignete Mechanismen vorhanden sind, um Fehler zu erkennen und mit der Zeit neues Wissen zu erwerben, kann dies nur dann in die klinische Anwendung übertragen werden, wenn der rechtliche Rahmen dies zulässt. Die derzeitigen Vorschriften in den USA und der Europäischen Union lassen nur abgeschlossene, deterministische Systeme zu, deren Parameter sich nicht mehr verändern dürfen. Glücklicherweise erkennen die Aufsichtsbehörden die Notwendigkeit eines modernen, *lebenszyklusorientierten Regulierungsansatzes* an. Im dritten Teil der Dissertation gehe ich auf diese Bemühungen ein, sowie auf andere nötige Aspekte der Entwicklung von Systemen, die während ihres Lebenszyklus weiterlernen.

Wir befinden uns endlich in einer Phase, in der medizinische Fachkräfte und Aufsichtsbehörden Deep Learning begrüßen, und in der die Zahl der kommerziell erhältlichen diagnostischen Radiologiesysteme schnell ansteigt. Dies eröffnet uns die Chance – und die Verantwortung – zu zeigen, dass diese Systeme während ihrer gesamten Lebensdauer sicher und effektiv sein können.

---

# Contents

---

<b>Introduction</b>	<b>10</b>
<b>I. Data Drift in Medical: Detection and Adaptation</b>	<b>13</b>
1. Domain Shift in Medical Imaging	14
2. Towards Automatic Quality Assurance: Detecting Silent Failures	16
3. Out-of-distribution Detection	18
3.1. The papers . . . . .	19
3.1.1. Self-supervised out-of-distribution detection for cardiac CMR segmentation . . . . .	19
3.1.2. Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation . . . . .	35
3.1.3. Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation . . . . .	50
3.2. Conclusions and outlook . . . . .	65
4. Epistemic Uncertainty Estimation	66
4.1. The paper: Improving robustness and calibration in ensembles with diversity regularization	67
4.2. Conclusions and outlook . . . . .	95
5. Assessing the Coherence of Model Predictions	96
5.1. The paper: Quality monitoring of federated Covid-19 lesion segmentation . . . . .	96
5.2. Conclusions and outlook . . . . .	105
6. Domain Adaptation and OOD Generalization	106
6.1. Training models to maintain stable performance across domains . . . . .	106
6.2. Adapting data to the training domain . . . . .	107
6.3. Conclusions and outlook . . . . .	108
<b>II. Continual Learning</b>	<b>109</b>
7. The Continual Learning Landscape	110
7.1. Key definitions . . . . .	110
7.2. Properly characterizing a continual setting . . . . .	111
7.2.1. Task identity and boundaries . . . . .	112
7.3. Quantifying continual performance . . . . .	113

---

7.4. Continual learning methods . . . . .	115
<b>8. Expansion Methods and Task-agnostic Learning</b>	<b>117</b>
8.1. The papers . . . . .	117
8.1.1. What is wrong with continual learning in medical image segmentation? . . . . .	117
8.1.2. Task-Agnostic continual hippocampus segmentation for smooth population shifts .	133
8.2. Conclusions and outlook . . . . .	148
<b>9. Learning Meaningful Representations</b>	<b>149</b>
9.1. Building an expressive latent space . . . . .	149
9.2. Continual learning in transformer architectures . . . . .	150
9.3. Conclusions and outlook . . . . .	150
<b>III. Towards Lifelong Learning in the Clinical Workflow</b>	<b>151</b>
<b>10. Practical Challenges Hindering Lifelong Learning</b>	<b>152</b>
<b>11. The Need for Unified Evaluation Standards</b>	<b>153</b>
11.1. The paper: Lifelong nnU-Net for standardized medical continual learning . . . . .	153
<b>12. Safe and Efficient Active Learning in Clinics</b>	<b>169</b>
12.1. The paper: Efficient 3D interactive segmentation with i3Deep . . . . .	171
12.2. Conclusions and outlook . . . . .	189
<b>13. Regulatory Landscape in the US and EU</b>	<b>190</b>
13.1. The road towards marketing SaMD in the USA and EU . . . . .	191
13.1.1. Receiving approval in the USA . . . . .	191
13.1.2. The Medical Device Regulation . . . . .	192
13.2. Planned regulations that embrace lifelong learning . . . . .	194
13.2.1. The FDA’s discussion paper and action plan . . . . .	194
13.2.2. The European AI Act . . . . .	196
13.2.3. Commonalities and a look into a future lifecycle regulatory approach . . . . .	197
13.3. Conclusions and outlook . . . . .	198
<b>IV. Summary and Future Perspectives</b>	<b>201</b>
<b>Acknowledgements</b>	<b>205</b>

---

# List of Publications

---

This list contains papers I co-authored during my doctoral studies on the topic of *Lifelong Learning in the Clinical Open World*. They are divided into the chapters where I describe or mention them and *emphasized* if included as part of the dissertation. Three manuscripts are currently under review (Pati et al., 2021; González et al., 2022c, 2023).

## Part I: Data Drift in Medical: Detection and Adaptation

### Chapter 1: Domain Shift in Medical Imaging

### Chapter 2: Towards Automatic Quality Assurance: Detecting Silent Failures

### Chapter 3: Out-of-distribution Detection

1. C. González and A. Mukhopadhyay. *Self-supervised out-of-distribution detection for cardiac cmr segmentation*. In *International Conference on Medical Imaging with Deep Learning*, pages 205–218. PMLR, 2021
2. C. González, K. Gotkowski, A. Bucher, R. Fischbach, I. Kaltenborn, and A. Mukhopadhyay. *Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation*. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 304–314. Springer, 2021
3. C. González, K. Gotkowski, M. Fuchs, A. Bucher, A. Dadras, R. Fischbach, I. J. Kaltenborn, and A. Mukhopadhyay. *Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation*. *Medical Image Analysis*, 82:102596, 2022a

### Chapter 4: Epistemic Uncertainty Estimation

1. H. A. Mehrtens, C. González, and A. Mukhopadhyay. *Improving robustness and calibration in ensembles with diversity regularization*. In *DAGM German Conference on Pattern Recognition*, pages 36–50. Springer, 2022
2. M. Fuchs, C. González, and A. Mukhopadhyay. *Practical uncertainty quantification for brain tumor segmentation*. In *International Conference on Medical Imaging with Deep Learning*, pages 407–422. PMLR, 2022

### Chapter 5: Assessing the Coherence of Model Predictions

1. C. González, C. L. Harder, A. Ranem, R. Fischbach, I. J. Kaltenborn, A. Dadras, A. M. Bucher, and A. Mukhopadhyay. *Quality monitoring of federated covid-19 lesion segmentation*. In *Bildverarbeitung für die Medizin 2022*, pages 38–43. Springer, 2022b

### Chapter 6: Domain Adaptation and OOD Generalization

1. A. Sanner, C. González, and A. Mukhopadhyay. *How reliable are out-of-distribution generalization methods for medical image segmentation?* In *DAGM German Conference on Pattern Recognition*, pages 604–617. Springer, 2021



- 
2. J. Kalkhof, C. González, and A. Mukhopadhyay. Disentanglement enables cross-domain hippocampus segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022

## **Part II: Continual Learning**

### **Chapter 7: The Continual Learning Landscape**

### **Chapter 8: Expansion Methods and Task-agnostic Learning**

1. C. González, N. Lemke, G. Sakas, and A. Mukhopadhyay. *What is wrong with continual learning in medical image segmentation?* arXiv preprint arXiv:2010.11008, 2023
2. C. González, A. Ranem, A. Othman, and A. Mukhopadhyay. *Task-agnostic continual hippocampus segmentation for smooth population shifts.* In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 108–118. Springer, 2022d

### **Chapter 9: Learning Meaningful Representations**

1. M. Memmel, C. González, and A. Mukhopadhyay. Adversarial continual learning for multi-domain hippocampal segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 35–45. Springer, 2021
2. A. Ranem, C. González, and A. Mukhopadhyay. Continual hippocampus segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3711–3720, 2022

## **Part III: Towards Lifelong Learning in the Clinical Workflow**

### **Chapter 10: Practical Challenges Hindering Lifelong Learning**

### **Chapter 11: The Need for Unified Evaluation Standards**

1. C. González, A. Ranem, D. P. dos Santos, A. Othman, and A. Mukhopadhyay. *Lifelong nnu-net: a framework for standardized medical continual learning.* 2022c

### **Chapter 12: Safe and Efficient Active Learning in Clinics**

1. K. Gotkowski, C. González, I. Kaltenborn, R. Fischbach, A. Bucher, and A. Mukhopadhyay. *i3deep: Efficient 3d interactive segmentation with the nnu-net.* In *International Conference on Medical Imaging with Deep Learning*, pages 441–456. PMLR, 2022
2. K. Gotkowski, C. González, A. Bucher, and A. Mukhopadhyay. M3d-cam: A pytorch library to generate 3d attention maps for medical deep learning. In *Bildverarbeitung für die Medizin 2021*, pages 217–222. Springer, 2021
3. S. Pati, S. P. Thakur, M. Bhalerao, S. Thermos, U. Baid, K. Gotkowski, C. González, O. Guley, I. E. Hamamci, S. Er, et al. Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*, 2021

### **Chapter 13: Regulatory Landscape in the US and EU**

---

# Introduction

---

Even the most optimistic machine learning researchers were awed at the milestones reached by deep learning during the past decade. Whether topping game leaderboards (Silver et al., 2016), generating realistic-looking images (Goodfellow et al., 2020) or simulating human speech (Brown et al., 2020), it seems there are few tasks that deep neural networks (DNNs) cannot solve if given enough training data in close-world environments. This has generated ambitious expectations on how our lives will change in the coming years.

In the healthcare sector, deep learning (DL) is increasingly viewed as an opportunity to counter rising costs, lack of personnel and global challenges such as pandemic preparedness (Makower et al., 2010; Parekh et al., 2020). DNNs have shown impressive performance in medical imaging tasks such as automatic lesion localization and tissue segmentation (Isensee et al., 2021) which are key for helping clinicians make downstream diagnostic decisions and select appropriate treatment plans. Leveraging these advances could pave the way for better healthcare even in middle and low-income countries and rural areas.

Unfortunately, these hopes are often crushed when *systems designed for close-world environments are set free in the dynamic open world*. There, it becomes evident that DNNs have difficulty extrapolating the learnt knowledge to a *shifted* – if only slightly so – data distribution (Hendrycks and Dietterich, 2018). Within the deep learning community, it has long been understood that models will learn the simplest solution for the training data and stay within the bounds of the training scenario (Beery et al., 2018). Yet this is rarely taken into account when products are designed and placed in the market. In fact, there are hundreds of medical decision support software products that are commercially available *today* and have not gone through external or multi-site validation (van Leeuwen et al., 2021; Wu et al., 2021).

Naturally, with the passage of time, the data used for training and the samples the model encounters during deployment grow further apart, a phenomenon commonly referred to as *data drift*. In the medical sector, image acquisition practices and disease patterns are constantly changing. This means that DNNs, which held the promise of *learning* as more data became available, actually *deteriorate* as time goes on. The great successes of DL were possible due to having access to large, heterogeneous datasets. One could therefore argue that although specific models may become outdated, the database can continue growing and future models will, in fact, improve when compared to their previous states. However, this neglects the fact that *data cannot always be centrally stored*. Particularly in the medical field, privacy regulations strongly limit the storage and transfer of patient data (European Commission, 2017b). It is, therefore, possible that certain training samples *will only be available for a limited duration*.

It should also be considered that retraining models with ever-growing datasets requires significant time and computational resources. Even *OpenAI's* hugely popular *Chat GPT* (Brown et al., 2020) cannot answer questions referring to events that took place after 2021. This limits where we can leverage DNNs and whether they can help us in time-critical situations where human resources are at full capacity. Consider, for instance, the Covid-19 pandemic. There was a lot of research, but disappointingly little

---

actual use of DL (Hu et al., 2020; Roberts et al., 2021). As we will later see, although there were several initiatives for releasing open-source code and annotated datasets of chest CT lesion segmentation (Liu et al., 2022a), models trained only with these cases did not generalize well enough to be used safely.

How, then, *can we best transfer the close-world success of DNNs to the dynamic clinical world and trust them with something as important as our well-being?* This is the core question I attempt to address in my thesis.

The first step comes with understanding what types of distribution shifts can occur in dynamic clinical settings. In this work, I focus on Computer Tomography (CT) and Magnetic Resonance Image (MRI) data. In Chapter 1, I give an overview of the different factors of change that can cause a gap in DNN performance. These result from changes in image acquisition – such as the use of a different scanner – the patient population or disease patterns. Unfortunately, we do not always know *which shifts* are introduced in the data distribution or *when* these take place. DNNs *fail silently*, so observing the magnitude of the outputs (often incorrectly interpreted as confidence) is not enough for deciding when a prediction is trustworthy. I go over techniques for detecting uncertain predictions in Chapters 2 through 5. In Chapter 6, I investigate to what extent we can mitigate the problem of data drift through out-of-distribution generalization and domain adaptation.

In Part II, I move on to how we can leverage new data through *continual learning*. The goal here is to adapt to changes in the environment without forgetting previous knowledge. We will see that if we sequentially fine-tune a model without mechanisms for information preservation, the performance increases for new data but substantially *decreases* for images from an earlier data distribution. From a technical perspective, the goal of continual learning is to train a model (or model ensemble) in such a fashion that the performance increases *for all seen data distributions*. Even in continual learning validations, close-world assumptions are often made, such as presuming that image precedence information is given and that distribution shifts happen suddenly at set intervals. I show how working under these assumptions can be problematic when applying continual learning to more realistic settings and suggest how we can leverage quality assurance mechanisms presented in previous chapters to avoid such relaxations.

In practice, continual learning techniques would enable *lifelong learning* where ML systems continue to train throughout their entire lifecycle. Coming back to the pandemic preparedness scenario from before, a good strategy would have been to start with a large database of pneumonia subjects, for which training could require several weeks, and incrementally update the model to better reflect the specific disease patterns that manifest in SARS-CoV-2 infections. If the predictions were assessed by appropriate quality control mechanisms, the model could at each time stage assist in the diagnosis of at least a fraction of the cases, alleviating the workload of primary care workers.

However, there are still several hurdles to overcome before we see systems in clinics that train throughout their lifecycle. The first is a lack of benchmarking standards for validating DNNs in continual scenarios. I address this problem in Chapter 11. Secondly, lifelong learning comprises close collaborations with clinicians on-site, who must collect appropriate training data and monitor the quality of the predictions, tasks which were traditionally the sole responsibility of the manufacturer. In Chapter 12, I show how active learning can enhance the collaboration between healthcare professionals and the ML system. Finally, as we will see in Chapter 13, the current regulatory framework simply does not allow for model updates to take place post-approval. This means that newly collected data cannot be used until there is a new product release and subsequent approval process, which can take several months. Fortunately, there *are* several initiatives from the responsible entities that are actively trying to establish *lifecycle regulatory protocols*. This change would not only allow for pre-determined modifications to take place post-approval, but also drive manufacturers to develop continuous quality assurance mechanisms.

---

A change in the regulatory landscape is urgently needed. There are currently several hundred ML-based medical software products in the European and North American market, yet these are neither permitted to adapt to changing data distributions nor made to sufficiently monitor their performance. Very few prospective studies have actually taken place, and several commercially available products do not even report external validation results (Wu et al., 2021). Software standards adopted by regulatory bodies are not set in isolation but instead defined by researchers (Azzouzi et al., 2022), so by formalizing and improving our evaluation practices, we can help define the forthcoming guidelines.

I firmly believe that the promises of DL in healthcare can be realized. But for this, we need to quickly move away from designing close-world systems and start considering the factors of change that our models will encounter in the dynamic open world. We also need to manage expectations for single methodologies and instead build pipelines that can accumulate knowledge over time; but also *fail safely* thanks to appropriate quality assurance mechanisms. This will lay the ground for commercial systems that clinical end users actually find helpful and that they enjoy cooperating with.

---

**Part I.**

**Data Drift in Medical:  
Detection and Adaptation**

---

# 1. Domain Shift in Medical Imaging

---

Humans have an innate capacity to extrapolate learned knowledge. We are taught traffic signs from colored pictures expecting that we will know when to cross the street; medical doctors learn about some conditions from textbooks and later recognize them in real patients. We tend to assume that deep learning models – which perform surprisingly well in some tasks – share this capability. Unfortunately, that is far from reality. DNNs will inevitably learn the simplest solution that maximizes their objective, a phenomenon referred to as *shortcut learning* (Beery et al., 2018). They will therefore only make meaningful predictions for inputs similar to those seen in the training data. Even slight perturbations in contrast or brightness can significantly decrease the performance of state-of-the-art computer vision architectures (Hendrycks and Dietterich, 2018).

This problem is exacerbated for medical imaging, where datasets are *smaller* due to stringent data privacy regulations and the cost of expert annotations; and *higher in dimension*. Popular computer vision datasets include *MNIST* (Deng, 2012) and *CIFAR-10* (Krizhevsky et al., 2020) with, respectively, 70 thousand  $28 \times 28$  and 60 thousand  $32 \times 23$  images. In contrast, the widely-used datasets from the *Medical Segmentation Decathlon* (Simpson et al., 2019) range from 30 (left atrium) to 750 (brain tumor) subjects and  $36 \times 50 \times 35$  (hippocampus) to  $512 \times 512 \times 482$  (liver) mean resolution.

Generalizing to even slight changes in the domain is therefore extremely difficult. And there are a lot of sources of domain shift that affect medical images and have been identified as problematic in the deployment of machine learning systems (Midya et al., 2018; Van Timmeren et al., 2020). Many are closely related to the acquisition and result from high variability between imaging protocols. For computer tomography (CT) and magnetic resonance imaging (MRI) data, these include:

- Scanner vendor (or even model)
- Acquisition settings, such as slice thickness, field strength (for MRIs), or tube current (for CTs)
- Factors that affect the position or geometry of the region of interest (ROI), such as field of view or the usage of coils
- Usage and timing of contrast agent
- Choice of the reconstruction algorithm
- Presence of image artifacts (e.g. due to movement, metallic foreign bodies, ghosting, or ringing)

In addition, there are factors related to the subject population that introduce geometric modifications and make it difficult to generalize to other geographical regions, including:

- Demographic factors (age, gender, heritage, etc.)
- Disease expression (phenotypes) and spread

- 
- Co-morbidity factors resulting from cultural or societal aspects

This makes *distribution* or *domain shift* a vital concern when training deep learning models for a number of medical imaging problems (Dou et al., 2019; Yan et al., 2019; Liu et al., 2022b). Yet we do not seem to grasp the scope of the problem and how much it hinders translating innovative research to clinics. Only very few medical ML solutions are validated in prospective clinical trials (Kelly et al., 2019; Nagendran et al., 2020), and some even lack proper external validation (Hu et al., 2020; Roberts et al., 2021). This causes ML models to display disappointing performance in real settings (Beede et al., 2020).

While strategies such as data augmentation and image harmonization may improve generalization, we do not always know which shifts we will encounter in the future. The data distribution changes over time, a phenomenon we often refer to as *concept* or *data drift* (Hoens et al., 2012), causing a gradual deterioration of machine learning model performance. The best way to counter this process is by adapting the model with new data samples, and we will explore strategies to do this in Part II.

But first, we will review how to **detect** (Chapters 2 through 5) when the model makes failed predictions with the help of automatic quality assurance mechanisms, and how to encourage **adaptation** to other (known) domains in Chapter 6.

---

## 2. Towards Automatic Quality Assurance: Detecting Silent Failures

---

One thing is safe with working with deep learning: no matter how robust our model is, there will always be cases for which it is simply not suitable for making a prediction. This only hinders the deployment of DNNs insofar as this translates to *silent failures* that we do not detect. If we are able to make an accurate assessment regarding the validity of each output, we can utilize the model for at least a fraction of test subjects.

While the commonly used *Softmax* function normalizes outputs so they add up to one, this should not be understood as a probability estimate. The typical way of training DNNs with backpropagation and an objective such as *Binary Cross Entropy* encourages outputs close to the discrete ground truth values, which will therefore tend to be over-confident (Hein et al., 2019). This brings us to the question of *how to reliably detect faulty predictions*, for which I believe there are three major strategies (illustrated in Figure 2.1) that all pursue the goal of *automatic quality assurance*.

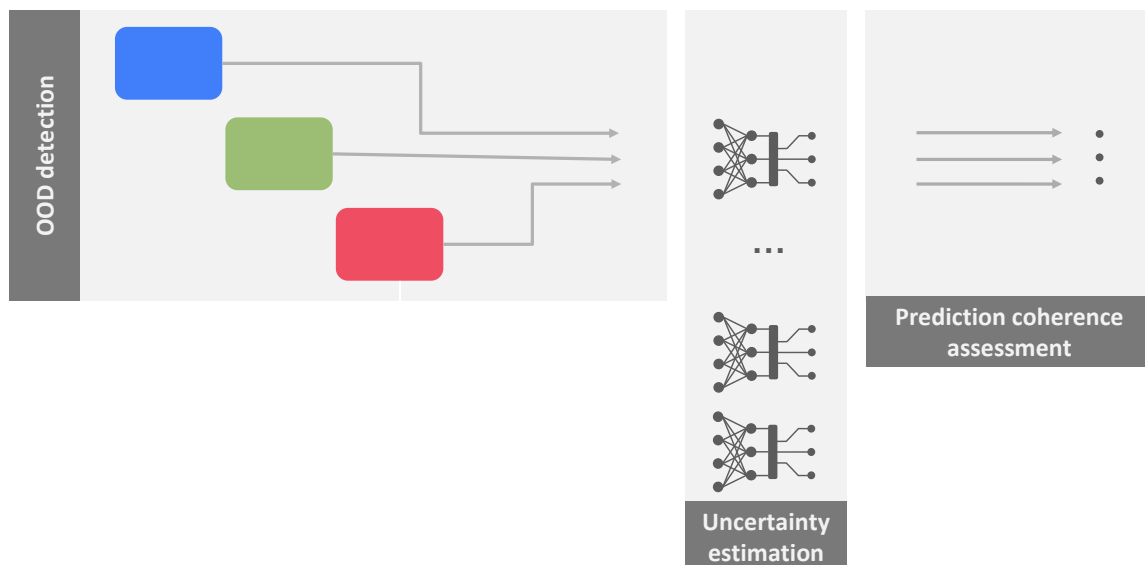


Figure 2.1.: Three mechanisms for automatic quality assurance: (1) out-of-distribution detection, (2) uncertainty estimation and (3) assessing the coherence of model predictions.

The first is *out-of-distribution detection*, where we identify inputs that the model was not trained to handle. This is the case when no meaningful prediction could ever be made for a given input, such as if a colon examination is erroneously fed to a model for lung segmentation; and also when the test case is so far from the training data that our model will not be able to generalize to it. The second would, for instance,



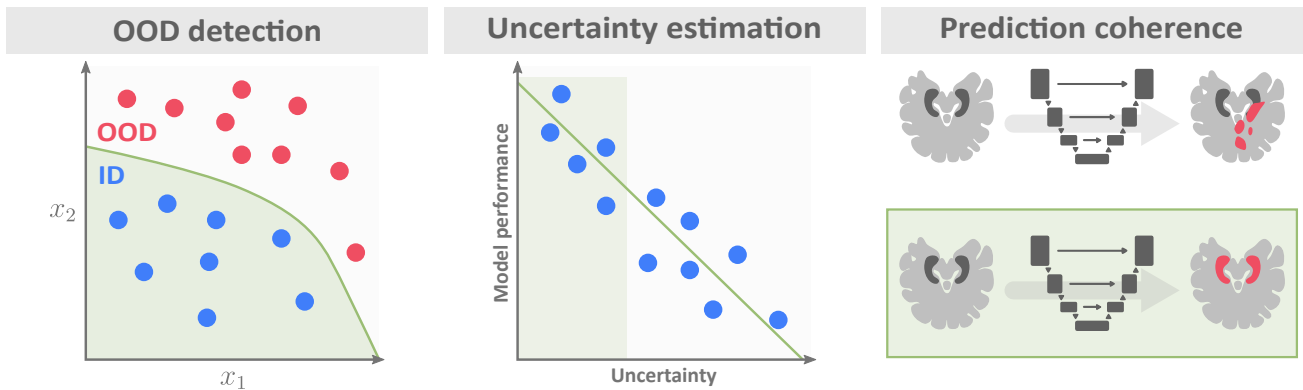


Figure 2.2.: Input space as deemed acceptable by different quality assurance mechanisms. OOD detection excludes samples far from the training distribution. Uncertainty estimation learns a function to calibrate the model, so only low-uncertainty predictions are considered. For certain tasks – such as image segmentation – observing properties of the outputs can help detect when these are erroneous.

take place when a scan from a child is given to a model trained only with adults. We typically aim to distinguish between **in-distribution (ID)** and **out-of-distribution (OOD)** cases (see Figure 2.2, left), and can further differentiate between *near* and *far-OOD* scenarios (Fort et al., 2021).

Secondly, we can employ techniques from *uncertainty estimation*. These are typically utilized for in-distribution data where the model would be expected to make a meaningful prediction; and are not deemed reliable in the presence of dataset shift (Ovadia et al., 2019). The objective is to obtain continuous uncertainty scores. Ideally, this would inversely correlate with model performance. We refer to this goal as *model calibration*, which is usually measured with the *Expected Calibration Error (ECE)* in its various forms<sup>1</sup>. For a specific confidence threshold, we can also evaluate the uncertainty method with the *coverage* on test data – quantifying model utilization – and performance of the predictions deemed to be confident enough.

Finally, there are certain tasks – particularly when we have domain knowledge such as medical imaging – where we can identify suspicious predictions simply by observing the model outputs. One example of this is *semantic segmentation*. By simply viewing the segmentation mask and quantifying certain aspects, such as the number of connected components and their geometric shapes, we can use domain knowledge to flag suspicious predictions. This is more challenging for anatomies such as lung lesions that take on diffuse shapes, but it is still a meaningful quality check (as we will see in Section 5.1). We find this direction to be particularly important for increasing trust in ML systems. If a radiologist is presented with a suggestion that has clear semantic inconsistencies, this will inevitably worsen their opinion of the system and – in the worse case – even DL in general.

Significant advances have been made in automatic quality assurance in recent years. However, we do not believe that any one single method would ever be sufficient. Instead, an array of strategies should be employed in tandem to monitor the performance of DNNs, and these should be integrated at various points of the clinical workflow.

<sup>1</sup>The ECE is very susceptible to hyperparameter settings such as bin size, so several variants have been proposed instead (Ashukha et al., 2019).

### 3. Out-of-distribution Detection

Out-of-distribution detection looks to identify samples far from the training distribution. We can roughly classify OOD detection methods into four categories, which we illustrate in Figure 3.1 along with their benefits and disadvantages.

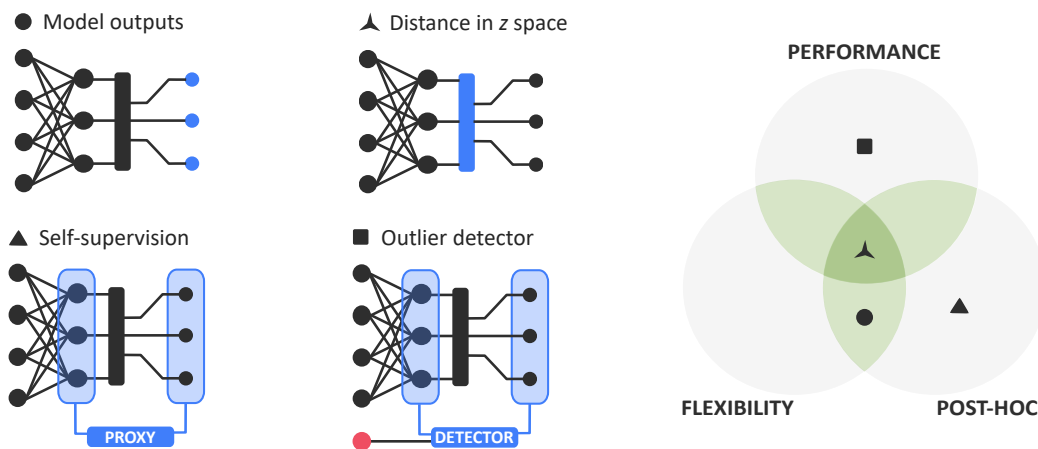


Figure 3.1.: Visual representation of four strategies for OOD detection. We can (1) quantify certain properties of the model outputs, such as the magnitude of the predicted class, (2) calculate how far the network features of test cases are from the training data, (3) monitor a self-supervision loss or (4) explicitly train an outlier detector with OOD samples. Ideally, the method should be effective at identifying OOD cases (*performance*) and applicable to any model architecture (*flexibility*) and to locked models without needing re-training (*post-hoc*).

The first strategy consists of **observing network outputs**. A simple baseline proposed by Hendrycks and Gimpel (2016) detects novelty by an equal distribution between logit values, i.e. how far these are from a one-hot encoding that would signal a confident prediction. Guo et al. (2017) discovered that using a temperature-scaled version of the *Softmax* function instead of the regular one leads to more accurate estimates, and Liang et al. (2018) built upon this with *ODIN*, which perturbs the inputs in an adversarial fashion to increase the distance between ID and OOD outputs. In a similar direction, Liu et al. (2020) employ an *Energy Scoring* function to identify out-of-distribution samples. The main advantage of output-based methods is that they are computationally inexpensive and work on a *purely post-hoc basis*. This means that OOD detection does not need to have been taken into account when designing the model architecture; and can instead be calculated even for pre-trained models.

If this is not a constraint, another possibility is to **explicitly train an outlier detector** that distinguishes between training ID and (real or simulated) OOD samples (Hendrycks et al., 2018; Lee et al., 2018a; Sabokrou et al., 2018; Vyas et al., 2018; Bevandić et al., 2019; Mohseni et al., 2020). This is a good option if one has some knowledge about the types of domain shifts that the model will encounter; and if

---

OOD detection is a main goal from the start. The main drawback is that the model architecture and loss function need to be modified to accommodate the additional OOD detection objective, which excludes the use of pre-trained models and may disrupt training. In addition, the model will only detect samples similar to OOD train samples, so there is no guarantee that unexpected shifts will be flagged.

A third category of methods leverages **self-supervision** to detect novelty. In self-supervised learning, we typically solve the *target task* that we are actually interested in together with a *proxy task*. The proxy task does not require manual annotations, so it allows us to leverage non-annotated data and encourages the model to learn more expressive representations (Asano et al., 2019). Besides improving the performance and robustness on the target task, self-supervised models have an additional advantage: as we can calculate the proxy loss during testing, we can use this as an OOD detection signal. The intuition is that for an OOD sample, the model will fail on the proxy task as much as it does for the target one. Proxy losses that have been utilized for this purpose include input reconstruction (Pidhorskyi et al., 2018; Xia et al., 2020), contrastive learning (Winkens et al., 2020; Wu and Goodman, 2020) and detection of transformations or rotation to the input (Golan and El-Yaniv, 2018; Hendrycks et al., 2019). Monitoring the proxy loss is undoubtedly a good practice when already working with a self-supervised model, and we will see how this idea can be exploited for cardiac CMR segmentation in Section 3.1.1. Nevertheless, it is doubtful whether augmenting a model with self-supervision solely for the goal of OOD detection is justifiable.

Finally, a fourth strategy looks at model features and flags inputs for which **test features diverge strongly from the training distribution**. Considering the way deep learning models learn and operate, we know that they will not produce reasonable outputs for activations in previous layers that are too far from those seen during training (Lee et al., 2018b). The challenge here is to obtain features that are small enough to estimate their distribution yet expressive enough to communicate a shift in the domain. In Section 3.1.2, we propose a method that uses this strategy for identifying low-quality segmentations. This method is also computationally efficient and works in a purely post-hoc manner, and we find it to work effectively at detecting OOD silent failures.

---

## 3.1. The papers

---

During my doctoral studies, I worked with two of the directions presented in the previous section: leveraging self-supervision and observing the distance to features in the training data. I will first describe the method we propose which combines uncertainty estimation with the test-time proxy loss value in Section 3.1.1. Afterward, I will outline our work flagging samples with a high Mahalanobis distance to the training distribution in a low-dimensional latent space, for which we published one conference and one journal paper (Sections 3.1.2 and 3.1.3).

### 3.1.1. Self-supervised out-of-distribution detection for cardiac CMR segmentation

We presented the work *Self-supervised out-of-distribution detection for cardiac CMR segmentation* (González and Mukhopadhyay, 2021) at the *Medical Imaging with Deep Learning (MIDL)* conference which took place from July 7<sup>th</sup> to 9<sup>th</sup>, 2021. The conference was initially planned for Lübeck, Germany, but instead took on a virtual format due to the Covid-19 pandemic. The paper was featured at the *RSIP Vision MIDL Daily* magazine, and later in the August 2021 issue of *Computer Vision News* (Anzarouth et al., 2021).

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Self-supervised out-of-distribution detection for cardiac CMR segmentation**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Self-supervised out-of-distribution detection for cardiac CMR segmentation" (González et al. 2021) was published as a full research paper at the "Medical Imaging with Deep Learning (MIDL)". It constitutes a joint work of Camila González and Anirban Mukhopadhyay.*

*This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].*

*As corresponding and leading author, C. González led the overall research design, literature review and writing process of the paper. The choice of methodological framework and experimental setup, and implementation of the code, were likewise done by C. González. The central implications of this work were derived by A. Mukhopadhyay as general advisor of this work, who also contributed with continuous feedback during all phases of the paper writing process. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Ich bin mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum: 01 / 09 / 2023                      01 / 12 / 2023

Unterschrift:                       

Camila González

Anirban Mukhopadhyay

# Self-supervised Out-of-distribution Detection for Cardiac CMR Segmentation

**Camila Gonzalez** CAMILA.GONZALEZ@GRIS.INFORMATIK.TU-DARMSTADT.DE and  
**Anirban Mukhopadhyay** ANIRBAN.MUKHOPADHYAY@GRIS.INFORMATIK.TU-DARMSTADT.DE  
*Technical University of Darmstadt, Karolinenpl. 5, 64289 Darmstadt, Germany*

## Abstract

The segmentation of cardiac structures in Cine Magnetic Resonance imaging (CMR) plays an important role in monitoring ventricular function, and many deep learning solutions have been introduced that successfully automate this task. Yet due to variabilities in the CMR acquisition process, images from different centers or acquisition protocols differ considerably. This causes deep learning models to *fail silently*. It is therefore crucial to identify out-of-distribution (OOD) samples for which the trained model is unsuitable. For models with a self-supervised proxy task, we propose a simple method to identify OOD samples that does not require adapting the model architecture or access to a separate OOD dataset during training. As the performance of self-supervised tasks can be assessed without ground truth information, it indicates during test time when a sample differs from the training distribution. The proposed method combines a voxel-wise uncertainty estimate with the self-supervision information. Our approach is validated across three CMR datasets and two different proxy tasks. We find that it is more effective at detecting OOD samples than state-of-the-art post-hoc OOD detection and uncertainty estimation approaches.

**Keywords:** out-of-distribution detection, self-supervision, distribution shift

## 1. Introduction

Despite significant advances in diagnostic deep learning research, the adoption of learning-based systems in clinical practice is very limited. One reason for this is the inability of models to generalize to out-of-distribution (OOD) samples in real clinical settings, coupled with their tendency to produce overconfident predictions. Most deep learning systems are evaluated on test data similar in distribution to that used for training. When testing takes place on data gathered from different pieces of equipment or with a different protocol, there is a noticeable drop in performance (Glocker et al., 2019).

Cardiac Cine Magnetic Resonance imaging (CMR), the gold-standard for non-invasive volumetric quantification, is particularly prone to shifts in image properties. The acquisition process requires breath-holding, which is difficult for patients with arrhythmias. As a consequence, variations in image quality are magnified (Oksuz et al., 2019; Ruijsink et al., 2020). Automatic cardiac segmentation that generalizes well to unseen manufacturers is still an open challenge (Bevandić et al., 2019; Yan et al., 2020). Clinical deployment of deep neural networks (DNNs) would comprise a two-step process where the plausibility of a model output being correct is considered alongside the prediction. Observing softmax outputs is not sufficient, as DNNs produce overconfident predictions for OOD data (Hein et al., 2019). Fig. 1 shows how the segmentation performance of a U-Net deteriorates silently on

OOD data. As OOD detection is a secondary goal, an ideal detector would integrate into any existing model and require no modifications in the architecture or training procedure.

In this work, we explore how self-supervision can help uncover OOD samples for the task of left ventricular blood pool segmentation, which is often utilized clinically to calculate parameters such as Ejection Fraction. DNNs only produce meaningful outputs for in-distribution (ID) data (Su et al., 2020). This manifests in a drop in performance for OOD samples and, accordingly, a higher loss between the predicted and target values. While the loss cannot be calculated during inference for supervised tasks, it *can* be for self-supervised tasks that derive target values from the input images. For self-supervised models, this opens the possibility to leverage the test-time performance as a signal for the identification of OOD samples without needing any manual annotations or OOD training data.

Our proposed method uses the value of the self-supervision loss in combination with post-hoc uncertainty estimation. While other works have used the self-supervision loss to detect OOD samples in classification tasks, we adopt this idea for medical image segmentation. Unlike current state-of-the-art, the proposed approach does not require a specific proxy task, or training the model with the explicit goal of OOD detection, and is therefore applicable to a wide array of self-supervised architectures. The proposed method outperforms state-of-the-art post-hoc approaches for OOD detection and uncertainty estimation across three CMR datasets and for two different proxy tasks: edge detection and contrastive learning. Our main contributions are: (A) the introduction of self-supervision as a lightweight OOD detector for cardiac CMR segmentation and (B) a thorough evaluation of OOD detection methods on CMR imaging for three datasets and two different self-supervised architectures.

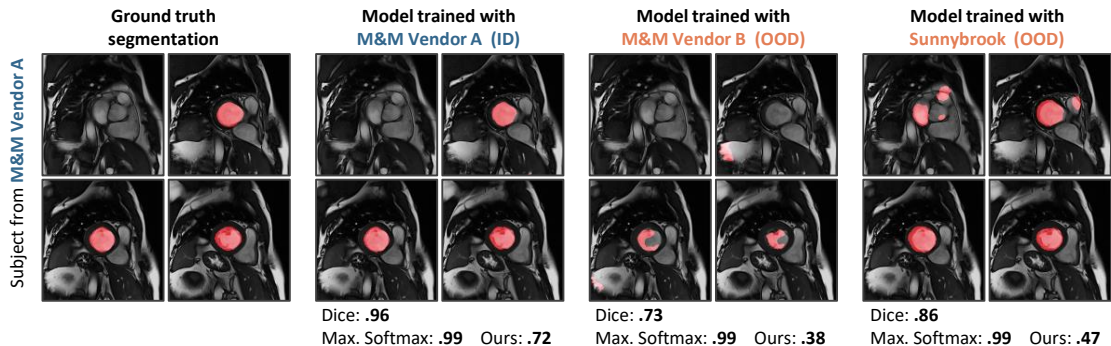


Figure 1: Distribution shift causes a deterioration on the left ventricular blood pool segmentation for a subject from the *Multi-Centre, Multi-Vendor and Multi-Disease (M&M) Vendor A* dataset, but traditional confidence quantification fails silently.

## 2. Related Work

In this section, we review relevant related work for self-supervision and OOD detection.

**Self-supervision** methods combine the training for the regular *target task* with a *proxy task*. Whereas the target task is usually supervised, the proxy task does not require manual

annotations, i.e. the target value can be derived from the input. For the sake of brevity we refer to [Asano et al. \(2019\)](#) and [Zhang et al. \(2019\)](#) for a detailed description of self-supervision in image segmentation.

In the field of **out-of-distribution detection**, several methods look at network outputs to detect novel samples. [Hendrycks and Gimpel \(2016\)](#) introduce the baseline of using the distribution of softmax values as an indicator for novelty. [Guo et al. \(2017\)](#) find temperature scaling to be an effective DNN calibration method. [Liang et al. \(2018\)](#) introduce the *ODIN* method, which extends temperature scaling by adding small adversarial-like perturbations to the inputs during inference which increase the separation between ID and OOD softmax values. [Lee et al. \(2018b\)](#) use the class-conditional distribution of neural activations to detect OOD samples. Other methods – that do not work in a post-hoc basis – use OOD data during training to explicitly train an outlier detector ([Hendrycks et al., 2018](#); [Lee et al., 2018a](#); [Mohseni et al., 2020](#); [Vyas et al., 2018](#); [Bevandić et al., 2019](#)). Related to the task of OOD detection is **uncertainty estimation**. Popular methods include Monte Carlo (MC) Dropout ([Gal and Ghahramani, 2016](#)) and Deep Ensembles ([Lakshminarayanan et al., 2017](#)). Several publications look at their effectiveness in the field of medical image segmentation, and find that ensembles are most reliable, though MC Dropout is also effective ([Jungo and Reyes, 2019](#); [Jungo et al., 2020](#); [Mehrtash et al., 2020](#)). Other methods have shown better performance in some cases, but require special training considerations ([Blundell et al., 2015](#); [Kohl et al., 2018](#); [Monteiro et al., 2020](#)).

Some research delves into **OOD detection in self-supervised models**. [Pidhorskyi et al. \(2018\)](#) use the reconstruction error of an autoencoder to assess novelty. [Winkens et al. \(2020\)](#) and [Wu and Goodman \(2020\)](#) augment classification networks with a contrastive learning term and estimate the density on different feature spaces. Similar to us, [Golan and El-Yaniv \(2018\)](#) train a multi-head model, where one head performs image classification and the second learns to detect image transformations, and calculate the novelty through the softmax outputs. [Hendrycks et al. \(2019\)](#) improve OOD detection by training a classifier with a proxy rotation estimation loss. For image segmentation, [Xia et al. \(2020\)](#) calculate the reconstruction error between the original image and a synthesized version.

Unlike other approaches, our proposed method does not require the use of a particular proxy task, and works entirely in a post-hoc manner. This ensures the applicability to a variety of deployed learning systems that include a self-supervised component. In terms of application we focus on semantic segmentation, and evaluate our method on datasets which solve the same semantic task (left ventricular blood pool segmentation) but differ in terms of acquisition vendor and center. Our research is, to our knowledge, the first to utilize self-supervision losses for OOD detection in medical image segmentation.

### 3. Methods

Consider a model  $\mathcal{F}$  trained with  $n$  samples  $\{x_i\}_{i=1}^n$ . The goal of OOD detection is to identify – during deployment – new samples that variate significantly from the training distribution. For this, a continuous *novelty* function  $\mathcal{N} : \mathcal{X} \rightarrow \mathbb{R}$  and a threshold  $\psi$  are defined so that  $x_i$  is classified as out-of-distribution if  $\mathcal{N}(x_i) \geq \psi$ . The expectation is that real-world OOD samples are flagged for which the model produces unreliable predictions. In this section, we describe our proposed method to detect OOD samples in a post-hoc

manner for models trained with a self-supervised proxy task. We start by introducing the two architectures we explore in this work, and then explain the process of OOD detection.

### 3.1. Self-supervised Learning

A task is said to be *self-supervised* if the target information is generated by the learning system. Increasingly, DNNs for semantic segmentation are being augmented with self-supervision (Wang et al., 2020; Pan et al., 2020) in order to leverage non-annotated data or shape the feature space. In this work, we explore **edge detection** and **contrastive learning**. These proxy tasks are well-suited to the segmentation of cardiac structures as they encourage learning geometrically-aware features that disregard image quality information (Chu et al., 2020; Winkens et al., 2020; Sahu et al., 2020). However, the novelty metric we introduce in Sec. 3.2 can be calculated for models trained with any self-supervised task.

**Contrastive learning** teaches the model to distinguish between different data points in the training set, while at the same time learning a semantically meaningful feature space that disregards certain transformations. Inspired by Winkens et al. (2020), we transform an original image  $x_i$  into  $\mathcal{T}(x_i) = \bar{x}_i$ . During training, we maximize the cosine similarity between  $x_i$  and  $\bar{x}_i$  in the feature space and minimize the similarity between  $x_i$  and a second image  $x_j$ . For function  $\mathcal{T}$ , we use implementations from the *TorchIO* library (version 0.17.46) (Pérez-García et al., 2020). We randomly apply *RescaleIntensity*, *RandomGamma*, *RandomMotion*, *RandomBiasField*, *RandomNoise* and *RandomBlur* operations, each with a probability of  $p = 0.5$ . Features  $z_i$  are extracted from the output of the encoder  $\mathcal{E}$ . Eq. 1 defines the contrastive loss  $\mathcal{L}_{ss}^C$ , and the architecture is displayed in Fig. 2 (left).

$$\mathcal{L}_{ss}^C(x_i, x_j) = \mathcal{L}_{sim}(\mathcal{E}(x_i), \mathcal{E}(x_j)) - \mathcal{L}_{sim}(\mathcal{E}(x_i), \mathcal{E}(\mathcal{T}(x_i))), \quad \mathcal{L}_{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\|_2 \cdot \|z_j\|_2} \quad (1)$$

The goal of **edge detection** is to extract a mask of edges  $\hat{h}_i$  from image  $x_i$ . We train a standard two-headed architecture consisting of a shared encoder  $\mathcal{E}$  and two decoders,  $\mathcal{G}$  for the segmentation task and  $\mathcal{H}$  for edge detection. Fig. 2 (right) outlines the proposed architecture. We train both heads with a combined loss of Dice ( $\mathcal{L}_{Dice}$ ) and binary cross entropy ( $\mathcal{L}_{BCE}$ ) weighted equally. To produce target masks  $h_i$  in a deterministic manner, we use the *Canny Edge* detector (Canny, 1986) of the *Scikit Learn* (Pedregosa et al., 2012) library (version 0.24.1) with lower and upper bounds of, respectively, 150 and 200. During inference, we treat the edge detection loss  $\mathcal{L}_{ss}^E$  (Eq. 2) as a component of our novelty metric.

$$\mathcal{L}_{ss}^E(x_i, h_i) = \mathcal{L}_{Dice}(\mathcal{H}(x_i), h_i) + \mathcal{L}_{BCE}(\mathcal{H}(x_i), h_i) \quad (2)$$

### 3.2. Novelty Estimation

For detecting OOD samples during inference we combine uncertainty estimates with the loss of the self-supervised proxy task. Uncertainty estimation produces good calibrations in ID data, but often fails in the presence of dataset shift (Ovadia et al., 2019). We expect dataset shift to manifest in an unusually large self-supervision loss (Su et al., 2020) that compensates for the decreased ability to detect uncertain cases of uncertainty estimation methods. By combining these two factors, we obtain a reliable detection signal.



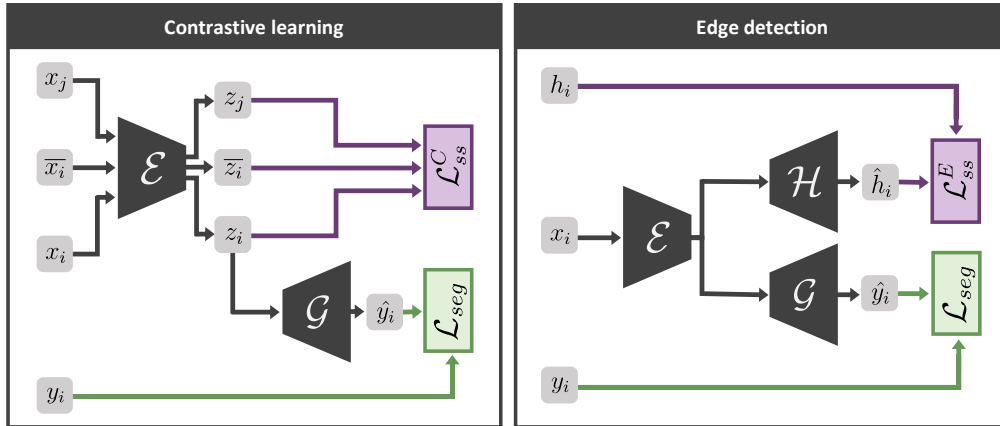


Figure 2: Two self-supervision architectures are explored in this work. Left: features are extracted for  $x_i$ ,  $\mathcal{T}(x_i) = \bar{x}_i$  and  $x_j$  to calculate a contrastive loss term. Right: network with an additional decoder head for the task of edge detection.

As we aim to find a flexible post-hoc method applicable to most learning-based systems, we explore two different types of uncertainty estimation. **MC Dropout** (Gal and Ghahramani, 2016) involves performing several forward passes with dropout during test time. The method can be applied to any model that uses dropout layers, which includes most modern architectures. **Deep Ensembles** – the practice of training several networks and averaging their predictions – have consistently shown the best performance in uncertainty estimation (Jungo et al., 2020; Mehrtash et al., 2020). They are also a straightforward way to improve prediction performance and therefore often used in practice. In the event that several trained models are present, we propose using this method as an uncertainty estimate.

During inference, the novelty of a test subject is assessed by combining the self-supervised loss  $\mathcal{L}_{ss}$  with uncertainty estimation. The  $\mathcal{L}_{ss}$  loss is calculated in the same way as during training. For the experiments performed in this work, either  $\mathcal{L}_{ss}^C(x_i, x_j)$  or  $\mathcal{L}_{ss}^E(x_i, h_i)$  are calculated depending on the model architecture. In the first case, we use a different subject from the same dataset as  $x_j$ . For 2D models, the loss for a test subject is the average across slices, as is also the case during training. As the uncertainty estimation component we take the voxel-wise standard deviation between model predictions, which is averaged over all voxels to produce a subject-level score. Different predictions are obtained by performing MC Dropout or, if ensembles are available, by making a prediction with each model. We define the proposed novelty function  $\mathcal{N}$  in Eq. 3, where  $K$  is the number of trained models or dropout forward passes and  $N$  is the number of voxels  $x_{i,j}$  in an image  $x_i$ .

$$\mathcal{N}(x_i) = \lambda \mathcal{L}_{ss}(\cdot) + \frac{1}{N} \sum_{j=1}^N \sqrt{\frac{1}{K} \sum_{k=1}^K (x_{i,j}^k - \mu_{i,j})^2}, \quad \mu_{i,j} = \frac{1}{K} \sum_{k=1}^K x_{i,j}^k \quad (3)$$

## 4. Experimental Setup and Results

We use three CMR datasets. The first two are part of the *Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation* (M&M) dataset (Campello and Lekadir, 2020) and contain healthy subjects as well as subjects with hypertrophic and dilated cardiomyopathies. We use the data for vendors *A* and *B*, for which ground truth segmentations are available. The images were acquired with *Siemens Avanto* and *Philips Achieva* scanners, respectively, at different centers. Each dataset contains 75 subjects. Lastly, we use the *Sunnybrook Cardiac Data* (Radau et al., 2009), acquired at a different center with a *General Electric Signa* scanner. The data consists of 45 scans from healthy as well as diseased subjects suffering from hypertrophy and heart failure. All images were acquired with 1.5T fields strength. We extract from each subject the segmented diastolic and systolic phase volumes.

We train a slice-by-slice U-Net with five encoding blocks based on the implementation by Pérez-García (2020). Images are center-cropped to  $256 \times 256$ . Each model is trained for 200 epochs with the *PyTorch Adam* optimizer. For the edge detection task, the encoder is shared and the decoder is replicated from the point with minimum spatial resolution. Refer to Appendix A for an overview of segmentation performance in ID and OOD data. Note that the results on the target task change slightly due to the incorporation of self-supervision.

We compare the proposed method against taking the inverse maximum softmax value (Hendrycks and Gimpel, 2016) (reported as **Max. Softmax**), temperature scaling (**Temp. Scaling**) (Guo et al., 2017) and the **ODIN** method (Liang et al., 2018); as well as against the corresponding uncertainty estimation (**MC Dropout** and **Ensemble**) and using only the self-supervised loss as a novelty estimate (**SS Loss**). When necessary, we average voxel-wise estimates to produce a volume-wise novelty score. We refer to our method variations using and not using ensembles as **Ours E** and **Ours**, respectively. We further specify in parenthesis whether the model learned a contrastive (C) or edge detection (E) task.

In turn, we consider each of the three datasets as ID and the other two as OOD. We divide the ID cases into three folds to perform cross-validation. For each cross-validation run, we train a model with the *ID train data* made out of two folds and evaluate it with the third fold, which is the *ID test data*. For OOD detection, we use one OOD dataset and the ID train data to select the best hyperparameters and evaluate the detection performance on the second OOD dataset and the ID test samples. We average the results of using each of the two OOD datasets for the evaluation, and report the mean and standard deviation of the three-fold cross-validation. Refer to Appendix D for a graphical illustration of our evaluation strategy. The following hyperparameters are tested:  $T \in \{1e1, 1e2, 1e3\}$  for temperature,  $\varepsilon \in \{1e-1, 1e-2, 1e-3\}$  for perturbation magnitude (ODIN),  $p \in \{0.3, 0.5, 0.7\}$  for dropout probabilities and  $\lambda \in \{1e0, 1e2, 1e4\}$  for weighting magnitudes.

We train ensembles with  $K = 3$  models and perform  $K = 30$  MC Dropout passes. We select the threshold  $\psi$  that achieves a 95% True Positive Rate (TPR) in the in-distribution train data, and flag samples as OOD when  $\mathcal{N}(x) \geq \psi$ . Reported are the Detection Error as defined by Liang et al. (2018) and the False Positive Rate (FPR) at 95% TPR.

### 4.1. Results for Contrastive Learning Models

We start by analyzing the results of OOD detection methods for the models trained with a contrastive learning loss component. Table 1 summarizes our findings. We see that for

Table 1: OOD Detection Error and FPR at 95% TPR for models trained with a contrastive learning loss term (lower is better). The mean and standard deviation are reported of testing with each OOD dataset and performing three-fold cross validation.

Method	M&M Vendor A		M&M Vendor B		Sunnybrook	
	Error	FPR	Error	FPR	Error	FPR
Max. Softmax	.48 ±.00	.93 ±.01	.51 ±.02	.90 ±.02	.53 ±.00	.91 ±.09
Temp. Scaling	.51 ±.01	.93 ±.01	.51 ±.02	.93 ±.01	.47 ±.01	.90 ±.02
ODIN	.43 ±.02	.84 ±.03	.49 ±.00	.87 ±.01	.51 ±.01	.87 ±.02
SS Loss (C)	<b>.33 ±.03</b>	.61 ±.04	.36 ±.11	.60 ±.17	.50 ±.04	.91 ±.02
MC Dropout	.45 ±.01	.85 ±.05	.38 ±.10	.72 ±.20	.21 ±.02	.23 ±.09
Ours (C)	<b>.33 ±.03</b>	<b>.60 ±.05</b>	<b>.33 ±.12</b>	<b>.58 ±.18</b>	<b>.19 ±.02</b>	<b>.19 ±.09</b>
Ensemble	.46 ±.02	.86 ±.01	.44 ±.03	.37 ±.08	<b>.26 ±.01</b>	.06 ±.02
Ours E (C)	<b>.32 ±.05</b>	<b>.49 ±.13</b>	<b>.26 ±.05</b>	<b>.17 ±.04</b>	.28 ±.01	<b>.05 ±.00</b>

all datasets, the popular temperature scaling and ODIN methods perform poorly. This may be due to the fact that both methods are developed for the classification task and not segmentation, where different voxels may be more or less significant for determining whether a sample is in-distribution. Our proposed method results in a lower detection error and FPR than all baselines both in cases where ensembles are available and when they are not. Only in dataset *Sunnybrook* does the ensemble alone achieve a lower detection error than the proposed method. As expected, considering the deviation between ensembles as an uncertainty estimation component leads to better results than applying MC Dropout. However, this method variation is only applicable if multiple models have been trained.

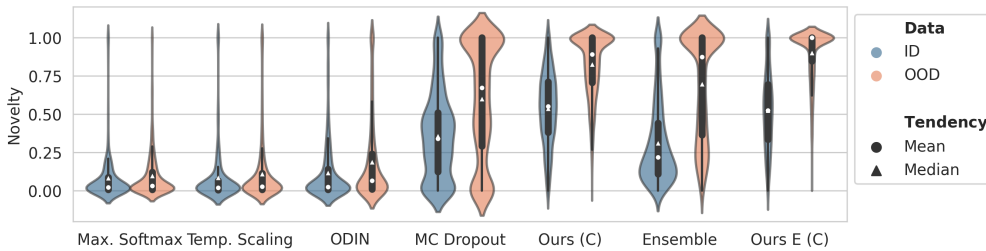


Figure 3: Distribution of novelty scores for contrastive learning models (lesser overlap is better). The scores for ID and OOD data are aggregated for all experiments and normalized to  $[0, 1]$  by taking the range of the ID training set.

Fig. 3 illustrates the ranges that different novelty scores occupy, normalized by taking the minimum and maximum novelty for ID train data, so that different methods are comparable. Ideally, novelty scores would cluster close to one (upper plot segment) for OOD data, and there would be a minimal overlap between ID and OOD scores. By observing the boxes

ranging from the first to the third quantiles we notice that the proposed method achieves the best separation between ID and OOD novelty scores in its two variations.

#### 4.2. Results for Architectures with Edge Detection

Table 2 compiles the results for models trained with an edge detection proxy task. Despite this being a very different task and self-supervision loss, the proposed method still performs best in all but one cases. However, the method shows its limitations for models trained with data from *M&M Vendor B*. This indicates that although our method is suited to any self-supervised task, some tasks may be more helpful than others.

Table 2: OOD Detection Error and FPR at 95% TPR ( $\pm$  standard deviation) for models trained with an edge-detection proxy task (lower is better).

Method	M&M Vendor A		M&M Vendor B		Sunnybrook	
	Error	FPR	Error	FPR	Error	FPR
Max. Softmax	.49 $\pm$ .00	.97 $\pm$ .00	.49 $\pm$ .01	.95 $\pm$ .04	.50 $\pm$ .00	.96 $\pm$ .01
Temp. Scaling	.51 $\pm$ .01	.87 $\pm$ .02	.51 $\pm$ .03	.91 $\pm$ .01	.48 $\pm$ .02	.92 $\pm$ .04
ODIN	.47 $\pm$ .02	.90 $\pm$ .01	.48 $\pm$ .03	.89 $\pm$ .02	.48 $\pm$ .01	.92 $\pm$ .00
SS Loss (E)	.33 $\pm$ .01	.66 $\pm$ .01	.55 $\pm$ .03	.99 $\pm$ .01	.29 $\pm$ .01	.53 $\pm$ .04
MC Dropout	.43 $\pm$ .04	.81 $\pm$ .04	<b>.44 <math>\pm</math> .06</b>	<b>.33 <math>\pm</math> .33</b>	<b>.28 <math>\pm</math> .03</b>	.28 $\pm$ .16
Ours (E)	<b>.32 <math>\pm</math> .01</b>	<b>.63 <math>\pm</math> .01</b>	<b>.44 <math>\pm</math> .05</b>	.81 $\pm$ .16	<b>.28 <math>\pm</math> .02</b>	<b>.25 <math>\pm</math> .14</b>
Ensemble	.39 $\pm$ .04	.68 $\pm$ .14	<b>.45 <math>\pm</math> .03</b>	.45 $\pm$ .02	.37 $\pm$ .13	.51 $\pm$ .49
Ours E (E)	<b>.32 <math>\pm</math> .01</b>	<b>.55 <math>\pm</math> .08</b>	<b>.45 <math>\pm</math> .03</b>	<b>.44 <math>\pm</math> .02</b>	<b>.25 <math>\pm</math> .01</b>	<b>.23 <math>\pm</math> .22</b>

## 5. Conclusion

Automatic segmentation of cardiac structures in CMR data could significantly alleviate the burden of clinicians. Competitive performance has been achieved by DNNs, but as long as these are susceptible to domain shift their applicability is limited. One way to approach this is by identifying OOD samples during deployment. For self-supervised models, combining the test-time value of the proxy loss with uncertainty estimation forms a reliable and lightweight novelty score. This finding is significant when considering the surge in popularity of self-supervision and introduces a further benefit of including a proxy term in DNN training. The proposed method can augment a wide array of learning-based systems, although for fully-supervised models incorporating a proxy task can have unintended effects in the target task. Future work should contemplate whether our results extend to other proxy tasks and anatomies. As it requires minimal overhead, we hope that monitoring the proxy loss during deployment becomes a widespread method for quality assurance.

## Acknowledgments

This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].

## Appendix A. Segmentation Performance of Trained Models

Table 3 showcases the Dice coefficient for left ventricular blood pool segmentation for models trained with two proxy tasks (contrastive and edge detection), as well as without any proxy task. In the diagonal, the results are displayed of testing each model with ID data.

Table 3: Mean Dice for models trained with a contrastive learning loss component (first row), edge detection (second row) and no self-supervised loss (third row). Reported are the mean and standard deviation of three cross-validation runs.

	Data	$\mathcal{F}$ trained with M&M Vendor A	$\mathcal{F}$ trained with M&M Vendor B	$\mathcal{F}$ trained with Sunnybrook
$\mathcal{L}_{ss}^C$	M&M Vendor A	<b>.85 ±.02</b>	.37 ±.05	.57 ±.02
	M&M Vendor B	.71 ±.01	<b>.87 ±.02</b>	.44 ±.10
	Sunnybrook	.57 ±.03	.14 ±.04	<b>.83 ±.02</b>
$\mathcal{L}_{ss}^E$	M&M Vendor A	<b>.83 ±.04</b>	.36 ±.05	.50 ±.05
	M&M Vendor B	.65 ±.02	<b>.86 ±.02</b>	.35 ±.15
	Sunnybrook	.60 ±.03	.09 ±.03	<b>.82 ±.01</b>
No $\mathcal{L}_{ss}$	M&M Vendor A	<b>.86 ±.02</b>	.42 ±.05	.60 ±.06
	M&M Vendor B	.71 ±.07	<b>.87 ±.06</b>	.36 ±.08
	Sunnybrook	.53 ±.02	.16 ±.08	<b>.80 ±.06</b>

## Appendix B. Novelty Distribution for Edge Detection Models

Fig. 4 displays the distribution of novelty scores for models with an edge detection proxy task. We see that the amount of overlap between ID and OOD data is more pronounced than for contrastive learning models (see Fig. 3). The variant of our method that uses ensembles (*Ours E (E)*) is the only approach that achieves a good separation.

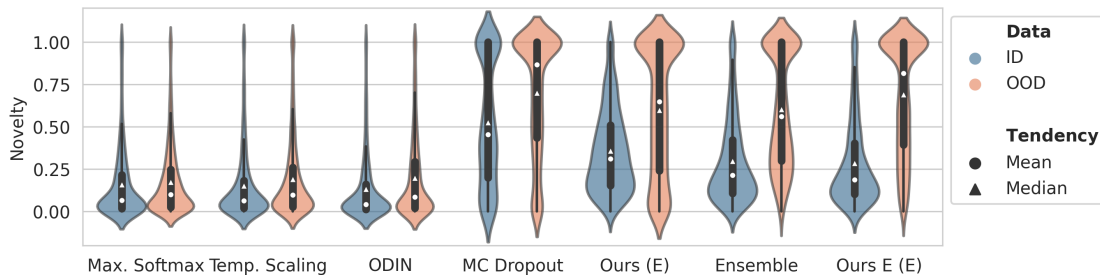


Figure 4: Distribution of novelty scores for models with an edge detection proxy task (lesser overlap is better). The novelty scores for ID and OOD data are aggregated for all experiments. The scores were normalized to  $[0, 1]$ .

### Appendix C. Generation of Target Data for Proxy Tasks

Fig. 5 displays exemplary data generated to train the proxy tasks explored in this work. The first column showcases slices from the *M&M Vendor B* dataset with overlaid ventricle blood pool segmentation (in red). The second column shows the same slices but with overlaid edge masks. Finally, the third column illustrates possible results of applying the transformation  $\mathcal{T}(x_i) = \bar{x}_i$ .

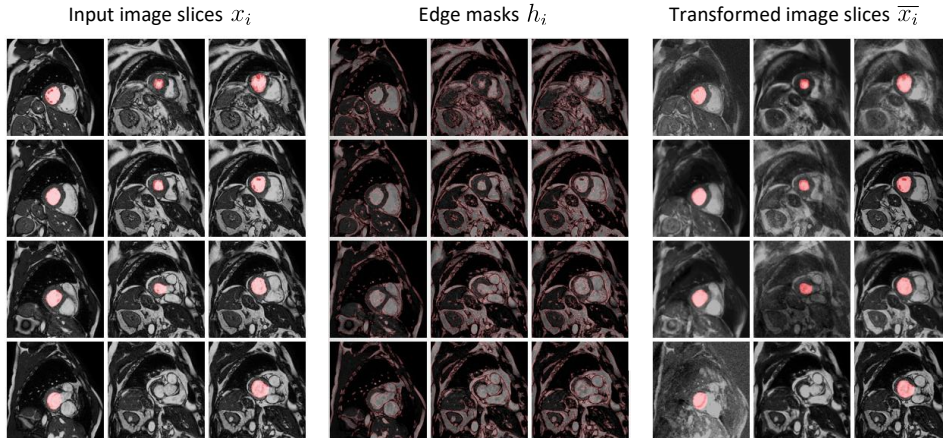


Figure 5: From left to right: image  $x_i$  with overlaid left ventricle blood pool segmentation ( $y_i$ ),  $x_i$  with overlaid edges  $h_i$  and transformed image  $\bar{x}_i$  with overlaid  $y_i$ .

### Appendix D. Evaluation Strategy

Fig. 6 graphically illustrates our evaluation setup with three datasets for one cross-validation run. In turn, each dataset is considered ID and is divided into *ID train* and *ID test* data. The ID train data is used to train the model, as well as to set hyperparameters alongside one OOD dataset. The detection performance is reported in the ID test data and the second OOD dataset. The results of using each OOD dataset for each purpose are averaged.

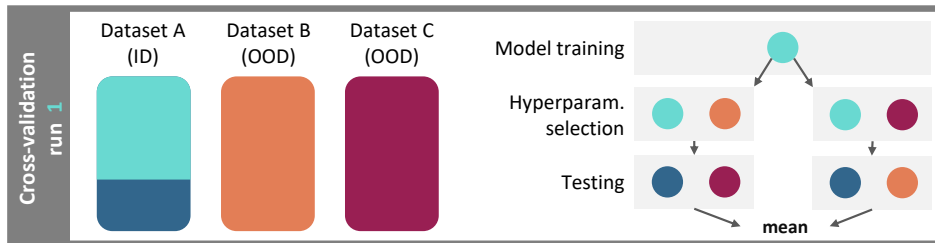


Figure 6: Graphical illustration of the evaluation strategy.

## References

- YM Asano, C Rupprecht, and A Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2019.
- Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *German Conference on Pattern Recognition*, pages 33–47. Springer, 2019.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.
- Victor M Campello and K Lekadir. Multi-centre multi-vendor & multi-disease cardiac image segmentation challenge (m&ms). In *Medical Image Computing and Computer Assisted Intervention*, 2020.
- John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 6:679–698, 1986.
- Jiajia Chu, Yajie Chen, Wei Zhou, Heshui Shi, Yukun Cao, Dandan Tu, Richu Jin, and Yongchao Xu. Pay more attention to discontinuity for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 166–175. Springer, 2020.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.
- Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.
- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

- Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.
- Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32:15663–15674, 2019.
- Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019.
- Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.
- Simon AA Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus H Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6965–6975, 2018.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.
- Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018a.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018b.
- Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.
- Alireza Mehrdash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.
- Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, pages 5216–5223, 2020.
- Miguel Monteiro, Loic Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12756–12767. Curran Associates, Inc., 2020.



- Ilkay Oksuz, Bram Ruijsink, Esther Puyol-Antón, James R Clough, Gastao Cruz, Aurelien Bustin, Claudia Prieto, Rene Botnar, Daniel Rueckert, Julia A Schnabel, et al. Automatic cnn-based detection of cardiac mr motion artefacts using k-space data augmentation and curriculum learning. *Medical image analysis*, 55:136–147, 2019.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 13991–14002. Curran Associates, Inc., 2019.
- Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020.
- Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.
- Fernando Pérez-García. fepegar/unet: PyTorch implementation of 2D and 3D U-Net, 03 2020. URL <https://doi.org/10.5281/zenodo.3697931>.
- Fernando Pérez-García, Rachel Sparks, and Sebastien Ourselin. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv:2003.04696 [cs, eess, stat]*, 03 2020. URL <http://arxiv.org/abs/2003.04696>. arXiv: 2003.04696.
- Stanislav Pidhorskyi, Ranya Almohten, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in neural information processing systems*, pages 6822–6833, 2018.
- Perry Radau, Yingli Lu, Kim Connelly, Paul Graham, Alexander Dick, and Graham Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009.
- Bram Ruijsink, Esther Puyol-Antón, Ilkay Oksuz, Matthew Sinclair, Wenjia Bai, Julia A Schnabel, Reza Razavi, and Andrew P King. Fully automated, quality-controlled cardiac analysis from cmr: validation and large-scale application to characterize cardiac function. *Cardiovascular Imaging*, 13(3):684–695, 2020.
- Manish Sahu, Ronja Strömsdörfer, Anirban Mukhopadhyay, and Stefan Zachow. Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 784–794. Springer, 2020.

- Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *European Conference on Computer Vision*, pages 645–666. Springer, 2020.
- Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 550–564, 2018.
- Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.
- Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.
- Mike Wu and Noah Goodman. A simple framework for uncertainty in contrastive learning. *arXiv preprint arXiv:2010.02038*, 2020.
- Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020.
- Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. Mri manufacturer shift and adaptation: Increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. *Radiology: Artificial Intelligence*, 2(4):e190195, 2020.
- Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. A survey of semi- and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, pages 1–30, 2019.

---

## Contribution and impact

Cardiac Cine Magnetic Resonance imaging (CMR) allows accurate non-invasive volumetric quantification of cardiac structures. However, automatic cardiac segmentation suffers from (1) MRIs varying notoriously across vendors and (2) the fact that the examination requires breath holding, which is particularly difficult for its target population and often results in motion artifacts.

Given the rise in popularity of self-supervised models for semantic segmentation (Bai et al., 2019; Araslanov and Roth, 2021), we explore whether monitoring the proxy loss during testing can help detect OOD images. As the proxy term only captures certain characteristics of the data, we combine this signal with uncertainty estimation, for which we test *Monte Carlo Dropout* and *Deep Ensembles*. To our knowledge, we were the first to look into this strategy for medical image segmentation.

Instead of limiting our study to a specific self-supervision component, we look into architectures with edge detection and contrastive learning proxy losses. Our goal is not to suggest one specific architecture that allows for OOD detection but instead to extract post-hoc novelty estimates from the trained model. Our empirical results show that this is a useful signal for identifying examinations acquired with a yet-unseen vendor.

## Discussion and limitations

The main limitation of the proposed method is that it can only be applied to models that contain a self-supervision module. I would advise against augmenting a model with a proxy term for the sole purpose of OOD detection. Besides, the self-supervised term will only supply information on certain aspects of the data. For instance, an edge detector may help detect a shift in MR vendor but generalize well to patients with yet-unseen conditions.

Regarding the scope of the study, it would be interesting to explore more ROIs and self-supervision strategies, as well as scenarios where only a portion of the training data is labeled. Several interesting works have emerged since the publication of this paper on using self-supervision, particularly contrastive learning, for OOD detection (Li et al., 2022; Qi et al., 2022; Wang et al., 2022) which shows this is a promising research direction. Contrastive learning has also been successfully applied to domain adaptation (Gu et al., 2022), confirming that it is a useful technique for learning meaningful features and increasing robustness.

### 3.1.2. Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation

Our publication *Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation* was presented as an oral at the 24<sup>th</sup> *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* which was held from September 27<sup>th</sup> to October 1<sup>st</sup>, 2021, as a virtual event. For the execution of this research, we cooperated closely with colleagues from the interventional radiology department at the University Hospital Frankfurt, who collected CT examinations from Covid-19 patients and delineated them with lesions characteristic of the infection. Thanks to the paper, I was awarded the *Young Scientist Award* by MICCAI Society, which recognizes up to five out of 1600+ submitted and 500+ accepted papers where the first author is an early-career researcher.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation" (González et al. 2021) was published as a full research paper at the "International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)". It constitutes a joint work of Camila González, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn and Anirban Mukhopadhyay.*

*This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].*

*As corresponding and leading author, C. González led the overall research design, management and writing process of the paper. C. González and K. Gotkowski contributed the literature review together. The experimental design and choice of methodological framework was done by C. González with supervision from A. Mukhopadhyay. A. Bucher, R. Fischbach and I. Kaltenborn were responsible for collecting and processing the in-house data and reviewed the manuscript from a clinical perspective. C. González and K. Gotkowski prepared and pre-processed the openly available data. C. González and K. Gotkowski worked on the implementation of the code, and C. González conducted the experiments. The methodology, results and conclusions were written by C. González. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor of this work, who contributed with continuous feedback during all phases of the paper writing process. All authors reviewed the final manuscript. All authors agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum:	09 / 01 / 2023	01 / 11 / 2023	01 / 10 / 2023	01 / 09 / 2023
Unterschrift:				
	Camila González	Karol Gotkowski	Andreas Bucher	Ricarda Fischbach
Datum:	01 / 12 / 2023	01 / 12 / 2023		
Unterschrift:				
	Isabel Kaltenborn	Anirban Mukhopadhyay		

# Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation<sup>\*</sup>

Camila Gonzalez<sup>1</sup>[0000-0002-4510-7309](✉), Karol Gotkowski<sup>1</sup>, Andreas Bucher<sup>2</sup>, Ricarda Fischbach<sup>2</sup>, Isabel Kaltenborn<sup>2</sup>, and Anirban Mukhopadhyay<sup>1</sup>

<sup>1</sup> Darmstadt University of Technology, Karolinenpl. 5, 64289 Darmstadt, Germany  
`camila.gonzalez@gris.informatik.tu-darmstadt.de`

<sup>2</sup> University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

**Abstract.** Automatic segmentation of lung lesions in computer tomography has the potential to ease the burden of clinicians during the Covid-19 pandemic. Yet predictive deep learning models are not trusted in the clinical routine due to *failing silently* in out-of-distribution (OOD) data. We propose a lightweight OOD detection method that exploits the Mahalanobis distance in the feature space. The proposed approach can be seamlessly integrated into state-of-the-art segmentation pipelines without requiring changes in model architecture or training procedure, and can therefore be used to assess the suitability of pre-trained models to new data. We validate our method with a patch-based nnU-Net architecture trained with a multi-institutional dataset and find that it effectively detects samples that the model segments incorrectly.

**Keywords:** out-of-distribution detection · uncertainty estimation · distribution shift.

## 1 Introduction

Automatic lung lesion segmentation in the clinical routine would significantly lessen the burden of radiologists, standardise quantification and staging of Covid-19 as well as open the way for a more effective utilisation of hospital resources. With this hope, several initiatives have gathered Computed Axial Tomography (CAT) scans and ground-truth annotations from expert thorax radiologists and released them to the public [6, 20, 23]. Experts have identified ground glass opacities (GGOs) and consolidations as characteristic of a pulmonary infection onset by the SARS-CoV-2 virus [24]. Deep learning models have shown good performance in segmenting these lesions. Particularly the fully-automatic *nnU-Net* framework [11] secured top spots (9 out of 10, including the first) in the leaderboard for the *Covid-19 Lung CT Lesion Segmentation Challenge* [7].

Such frameworks would ideally be utilised in the clinical practice. However, deep learning models are known to fail for data that considerably diverges from

---

<sup>\*</sup> Supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].

the training distribution. CAT scans are particularly prone to this *domain shift* problem [4]. The data showcased in the challenge is multi-centre and diverse in terms of patient group and acquisition protocol. A model trained with it would be presumed to produce good predictions for a wide spectrum of institutions. Yet when we evaluate a nnU-Net model on three other datasets, we notice a considerable drop in segmentation quality (see Fig. 1 (a)). Lung lesions do not manifest in large connected components (see Fig 4), so it is not trivial for a novice radiologist to identify an incorrect segmentation.

Clinicians can still leverage models trained with large amounts of heterogeneous data, but only alongside a process that identifies when the model is unsuitable for a new data sample. Widely-used segmentation frameworks *are not designed with OOD detection in mind*, and so a method is needed that reliably identifies OOD samples post-training while requiring minimal intervention.

Several strategies have shown good OOD detection performance in classification models. Hendrycks and Gimpel [8] propose using the maximum softmax output as an OOD detection baseline. Guo et al. [5] find that replacing the regular softmax function with a *temperature-scaled* variant produces truer estimates. This can be complemented by adding perturbations to the network inputs [19]. Other methods [10, 17] instead look at the KL divergence of softmaxed outputs from the uniform distribution. Some approaches use OOD data during training to explicitly train an outlier detector [1, 9, 17]. Bayesian-inspired techniques can also be used for outlier detection. Commonly-used are Monte Carlo Dropout [3] and Deep Ensembles [16]. These have shown promising results in the field of medical image segmentation [12, 13, 21]. Approaches that modify the architecture or training procedure have shown better performance in some cases, but their applicability to widely-used segmentation frameworks is limited [2, 15, 22].

We propose a method for OOD detection that is lightweight and seamlessly integrates into complex segmentation frameworks. Inspired by the work of Lee et al. [18], our approach estimates a multivariate Gaussian distribution from in-distribution (ID) training samples and utilises the Mahalanobis distance as a measure of uncertainty during inference. We compute the distance in a low-dimensional feature space, and down-sample it further to ensure a computationally inexpensive calculation. We validate our method on a patch-based 3D nnU-Net trained with multi-centre data from the *Covid-19 Lung CT Lesion Segmentation Challenge*. Our evaluation shows that the proposed method can effectively identify OOD samples for which the model produces faulty segmentations, and provides good model calibration estimates. Our contributions are:

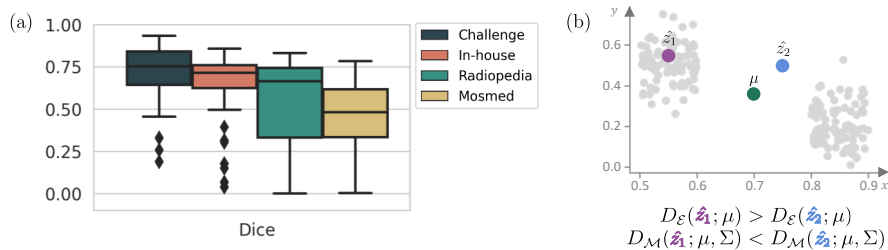
- The introduction of a lightweight, flexible method for OOD detection that can be integrated into any segmentation framework.
- An extension of the nnU-Net framework to provide clinically-relevant uncertainty estimates.

## 2 Materials and Methods

We start by summarising the particularities of the nnU-Net framework in Sec. 2.1. In Sec. 2.2, we outline our proposed method for OOD detection, which follows a *three-step process*: (1) estimation of a Gaussian distribution from training features (2) extraction of uncertainty masks for test images and (3) calculation of subject-level uncertainty scores.

### 2.1 Patch-based nnU-Net

The nnU-Net framework is a standardised baseline for medical image segmentation [11]. Without deviating from traditional U-Net architectures [26], it has won several grand challenges by automatically customising the architecture and training configuration to the data at hand [7]. The framework also performs pre- and post-processing steps, such as adapting voxel spacing and contrast normalisation, during both training and inference. In this work we utilise the patch-based full-resolution variant, which is recommended for most applications [11], but our method can be integrated into any other architecture. For the patch-based architecture, training images are first divided into overlapping patches with a sliding window approach, resulting in  $N$  patches  $\{x_i\}_{i=1}^N$ . Predictions for each patch are multiplied by a filtering operation that weights centre-voxels more heavily, and then aggregated into an output mask with the dimensions of the original image.



**Fig. 1.** (a) Dice coefficient of a model trained with *Challenge* data, evaluated with ID (*Challenge*) test data as well as on three other datasets. (b) The euclidean distance  $D_E$  does not recognize that  $\hat{z}_1$  (purple marker) is closer than  $\hat{z}_2$  (blue marker) to the distribution of training samples (gray markers), with mean  $\mu$  (green marker) and covariance  $\Sigma$ . This difference intensifies in high-dimensional spaces, where it is common for regions close to the mean to be underrepresented.

### 2.2 Estimation of a subject-level uncertainty score

We are interested in capturing *epistemic uncertainty*, which arises from a lack of knowledge about the data-generating process. Quantifying it for image *regions*

instead of region boundaries is challenging, particularly for OOD data [14]. One computationally inexpensive way to assess epistemic uncertainty is to calculate the distance between training and testing activations in a low-dimensional feature space. As a model is unlikely to produce reasonable outputs for features far from any seen during training, this is a reliable signal for bad model performance [18]. Model activations have covariance and the activations of typical input images do not necessarily resemble the mean [27], so the euclidean distance is not appropriate to identify unusual activation patterns; a problem that exacerbates in high-dimensional spaces. The Mahalanobis distance  $D_{\mathcal{M}}$  rescales samples into a space without covariance, supplying a more effective way to identify typical patterns in deep model features. Fig. 1 (b) illustrates a situation where the euclidean distance assumes that  $\hat{z}_2$  is closer to the training distribution than  $\hat{z}_1$ , when  $\hat{z}_2$  is highly unusual and  $\hat{z}_1$  is a probable sample.

In the following we describe the steps we perform to extract a subject-level uncertainty value. Note that only one forward pass is necessary for each image, keeping the computational overhead to a minimum.

**Estimation of the training distribution:** We start by estimating a multivariate Gaussian  $\mathcal{N}(\mu, \Sigma)$  over model features. For all training inputs  $\{x_i\}_{i=1}^N$ , features  $\mathcal{F}(x_i) = z_i$  are extracted from the encoder  $\mathcal{F}$  of the pre-trained model. For modern segmentation networks, the dimensionality of the extracted features  $z_i$  is too large to calculate the covariance  $\Sigma$  in an acceptable time frame. We thus project the latent space into a lower subspace by average pooling. Finally, we flatten this subspace and estimate the empirical mean  $\mu$  and covariance  $\Sigma$ .

$$\mu = \frac{1}{N} \sum_{i=1}^N \hat{z}_i, \quad \Sigma = \frac{1}{N} \sum_{i=1}^N (\hat{z}_i - \mu)(\hat{z}_i - \mu)^T \quad (1)$$

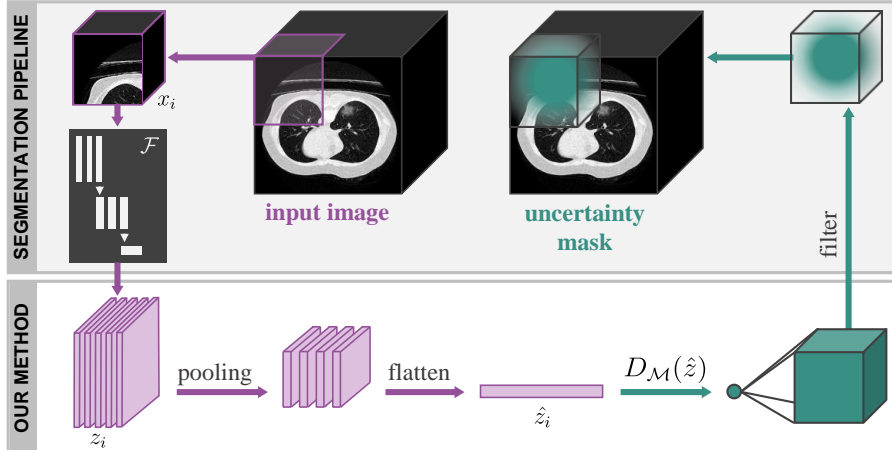
**Extraction of uncertainty masks:** During inference, we estimate an uncertainty mask for a subject following the process outlined in Fig. 2. For each patch  $x_i$ , features are extracted and projected into  $\hat{z}_i$ . Next, the Mahalanobis distance (Eq. 2) to the Gaussian distribution estimated in the previous step is calculated.

$$D_{\mathcal{M}}(\hat{z}_i; \mu, \Sigma) = (\hat{z}_i - \mu)^T \Sigma^{-1} (\hat{z}_i - \mu) \quad (2)$$

Each distance is a point estimate for the corresponding model input. These are aggregated in a similar fashion to how network outputs are combined to form a prediction mask. Following the example of the patch-based nnU-Net, a zero-filled tensor is initialised with the dimensionality of the original image. After assessing the distance for a patch, the value is replicated to the specified patch size and a filtering operation is applied to weight centre voxels more heavily. Finally, patch-level uncertainties are aggregated to an image-level mask.

**Subject-level uncertainty:** The process described above produces an uncertainty mask with the dimensionality of the CAT scan. In order to effectively





**Fig. 2.** Extracting an uncertainty mask based on the Mahalanobis distance  $D_{\mathcal{M}}$  for an image during inference, in combination with a patch-based nnU-Net architecture.

identify highly uncertain samples, we aggregate these into a subject-level uncertainty  $\mathcal{U}$  by averaging over all voxels. We then normalise uncertainties between the minimum and doubled maximum values represented in an ID validation set – which we assume to be available during training – to ensure  $\mathcal{U} \in [0, 1]$ .

### 3 Experimental Setup

We work with a total of four datasets for segmentation of Covid-19-related findings. The *Challenge* dataset [6] contains chest CAT scans for patients with a confirmed SARS-CoV-2 infection from an array of institutions. The data is heterogeneous in terms of age, gender and disease severity. We use the 199 cases made available under the *Covid Segmentation Grand Challenge*, which we randomly divide into 160 cases to train the model, 4 validation and 35 test cases.

We evaluate our method with two publicly available datasets and an in-house one. The public datasets encompass cases for patients with and without confirmed infections. *Mosmed* [23] contains fifty cases and the *Radiopedia* dataset [20], a further twenty. Finally, we utilise an in-house dataset consisting of fifty patients who were tested positive for SARS-CoV-2 with an RT PCR test. All fifty scans were reviewed for diagnostic image quality. The annotations for the in-house data were performed slice-by-slice by two independent readers trained in the delineation of GGOs and pulmonary consolidations. Central vascular structures and central bronchial structures were excluded from all segmentations. All delineations were reviewed by an expert radiologist reader. For the public datasets, the segmentation process is outlined in the corresponding publications.

With the *Challenge* data, we train a patch-based nnU-Net [11] on a *Tesla T4* GPU. Our configuration has a patch size of  $[28, 256, 256]$ , and adjacent patches

overlap by half that size. To reduce the dimensionality of the feature space, we apply average pooling with a kernel size of  $(2, 2, 2)$  and stride  $(2, 2, 2)$  until the dimensionality falls below  $1e4$  elements. With the *Scikit Learn* library (version 0.24) [25], calculating  $\Sigma$  requires 85 seconds for  $1e5$  samples. Our code is available under [github.com/MECLabTUDA/Lifelong-nnUNet](https://github.com/MECLabTUDA/Lifelong-nnUNet).

We compare our approach to state-of-the-art techniques to assess uncertainty information by performing inference on a trained model. *Max. Softmax* consists of taking the maximum softmax output [8]. *Temp. Scaling* performs temperature scaling on the outputs before applying the softmax operation [5], for which we test three different temperatures  $T = \{10, 100, 1000\}$ . *KL from Uniform* computes the KL divergence from an uniform distribution [10]. Note that all three methods output a *confidence* score (higher is more certain), which we invert to obtain an *uncertainty* estimate (lower is more certain). Finally, *MC Dropout* consists of doing several forward passes whilst activating the Dropout layers that would usually be dormant during inference. We perform 10 forward passes and report the standard deviation between outputs as an uncertainty score. For all methods, we calculate a subject-level metric by averaging uncertainty masks, and normalise the uncertainty range between the minimum and doubled maximum uncertainty represented in ID validation data.

## 4 Results

We start this section by analysing the performance of the proposed method in detecting samples that vary significantly from the training distribution. We then examine how well the model estimates segmentation performance. Lastly, we qualitatively evaluate our method for ID and OOD examples.

**OOD detection:** We first assess how effective our method is at identifying samples that are not ID (*Challenge* data). Due to the heterogeneity of the *Challenge* dataset, in practice data from an array of institutions would be considered ID. However, for our evaluation datasets there is a drop in performance which should manifest in higher uncertainty estimates. As is common practice in OOD detection [19], we find the uncertainty boundary that achieves a 95% true positive rate (TPR) on the ID validation set, where a *true positive* is a sample correctly identified as ID. We report for the ID test data and all OOD data the false positive rate (FPR) and  $Detection\ Error = 0.5(1 - TPR) + 0.5\ FPR$  at 95% TPR. Tab. 1 summarizes our findings. All methods that utilise the network outputs after one forward pass have a high detection error and FPR, while the MC Dropout approach manages to identify more OOD samples. Our proposed method displays the lowest FPR and detection error.

**Segmentation performance:** While the detection of OOD samples is a first step in assessing the suitability of a model, an ideal uncertainty metric would inversely correlate with model performance, informing the user of the likely quality of a prediction without requiring manual annotations. For this we calculate the *Expected Segmentation Calibration Error* (ESCE). Inspired by Guo

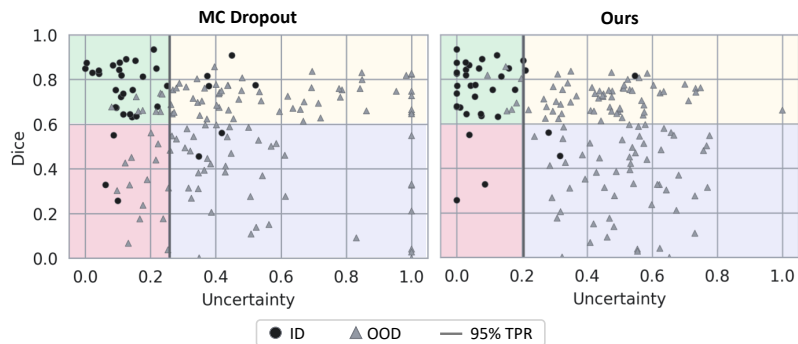
**Table 1.** Detection Error (lower is better) and FPR (lower is better) for the boundary of 95% TPR, ESCE (lower is better) and (mean±sd) Dice (higher is better) for subjects with an uncertainty below the 95% TPR boundary. The results are reported for ID test data and all OOD samples.

Method	Det. Error	FPR	ESCE	Dice
Max. Softmax [8]	0.334	0.583	0.319	0.582 ±0.223
Temp. Scaling $T = 10$ [5]	0.508	0.758	0.407	0.601 ±0.233
Temp. Scaling $T = 100$ [5]	0.361	0.550	0.408	0.589 ±0.233
Temp. Scaling $T = 1000$ [5]	0.500	1.000	0.408	0.592 ±0.233
KL from Uniform [10]	0.415	0.717	0.288	0.600 ±0.215
MC Dropout [3]	0.177	0.183	0.215	0.614 ±0.234
<b>Ours</b>	<b>0.082</b>	<b>0.050</b>	<b>0.125</b>	<b>0.744 ±0.143</b>

et al. [5], we divide the  $N$  test scans into  $M = 10$  interval bins  $B_m$  according to their normalised uncertainty. Over all bins, the absolute difference is added between average Dice ( $Dice(B_m)$ ) and inverse average uncertainty ( $1 - \mathcal{U}(B_m)$ ) for samples in the bin, weighted by the number of samples.

$$ESCE = \sum_{m=1}^M \frac{|B_m|}{N} |Dice(B_m) - (1 - \mathcal{U}(B_m))| \quad (3)$$

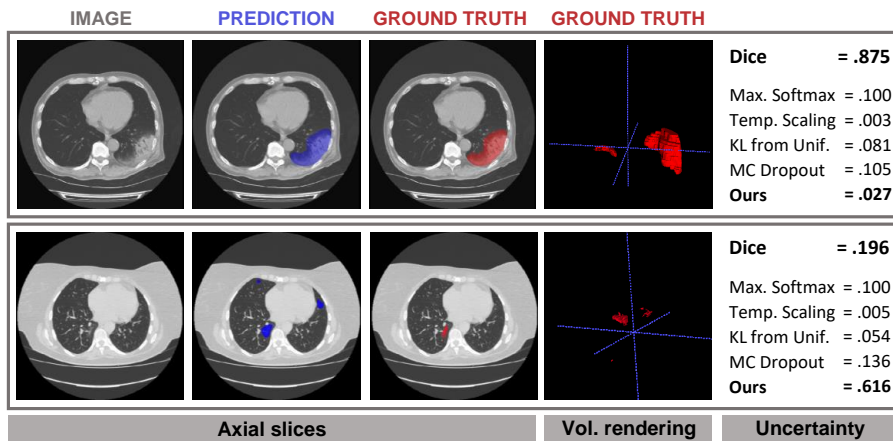
The results are reported in Tab. 1 (forth column). Our proposed approach shows the lowest  $ESCE$  at 0.125. The average Dice of admitted samples (fifth column) lies at 0.744, which is consistent with the ID expected performance of the model (see Fig. 1 (a)).



**Fig. 3.** Dice coefficient against normalised uncertainty for OOD (gray triangles) and test ID (black circles) samples. The vertical gray line marks the boundary of 95% TPR for ID validation data. To the right of this line, samples are classified as OOD. The lower left (red) quadrant is clinically most relevant. Unlike MC Dropout, our method does not fail silently by assigning low uncertainties to low-Dice samples.

Fig. 3 depicts the Dice coefficient plotted against the uncertainty for our proposed approach and MC Dropout, which has the second lowest calibration error. Relevant for a safe use of the model in clinical practice is the lower left (red) quadrant, where *silent failures* are located. Whereas MC Dropout fails to identify OOD samples with faulty predicted segmentations, our proposed method assigns these cases high uncertainty estimates. The only OOD samples that fall below the 95% TPR uncertainty boundary for our method have Dice scores over 0.6 (upper left quadrant with green background). However, our method shows room for improvement in the upper right (yellow) quadrant. Here, OOD samples for which the model produces good predictions are estimated to have a high uncertainty. An ideal calibration would place all samples in the upper left (green) and lower right (blue) quadrants.

**Qualitative evaluation:** Fig. 4 depicts two example images alongside corresponding ground truths and predictions. The top row shows a example from the *Challenge* dataset for which the model produces an adequate segmentation. The bottom contains a scan from the *Mosmed* dataset. The model oversegments the lesion at the middle left lobe and incorrectly marks two additional regions at the left and right superior lobes. Only our proposed method signals a possible error in the lower row with a high uncertainty, while producing a low uncertainty estimate for the upper row.



**Fig. 4.** Upper row: a good prediction. Lower row: a prediction for an OOD sample where two lesions are erroneously segmented in the superior lung lobes. Despite the considerable differences to the ground truth, these errors are not directly noticeable for the inexpert observer, as GGOs can manifest in superior lobes [24].

## 5 Conclusion

Increasingly, institutions are taking part in initiatives to gather large amounts of annotated, heterogeneous data and release it to the public. This could potentially alleviate the work burden of medical practitioners by allowing the training of robust segmentation models. Open-source end-to-end frameworks contribute to this process. But regardless of the variety of the training data, it is necessary to assess whether a model is well-suited to new samples. This is particularly true when it is not trivial to identify a faulty output, such as for the segmentation of SARS-CoV-2 lung lesions. There is currently a disconnect between methods for OOD detection, which often require special training or architectural considerations, and widely-used segmentation frameworks. We find that calculating the Mahalanobis distance to features in a low-dimensional subspace is a lightweight and flexible way to signal when a model prediction should not be trusted. Future work should explore how to better identify high-quality predictions and evaluate the methods considered in this work on other segmentation models. For now, our work increases clinicians' trust while translating trained neural networks from challenge participation to real clinics.

## References

1. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Simultaneous semantic segmentation and outlier detection in presence of domain shift. In: German Conference on Pattern Recognition. pp. 33–47. Springer (2019)
2. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning. pp. 1613–1622. PMLR (2015)
3. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. pp. 1050–1059. PMLR (2016)
4. Glocker, B., Robinson, R., Castro, D.C., Dou, Q., Konukoglu, E.: Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. arXiv preprint arXiv:1910.04597 (2019)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
6. Harmon, S.A., Sanford, T.H., Xu, S., Turkbey, E.B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al.: Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature communications* **11**(1), 1–7 (2020). <https://doi.org/https://doi.org/10.1038/s41467-020-17971-2>
7. Henderson, E.: Leading pediatric hospital reveals top ai models in covid-19 grand challenge. <https://www.news-medical.net/news/20210112/Leading-pediatric-hospital-reveals-top-AI-models-in-COVID-19-Grand-Challenge.aspx>, accessed: 2021-02-28
8. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
9. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations (2018)
10. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems* **32**, 15663–15674 (2019)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**(2), 203–211 (2021)
12. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience* **14**, 282 (2020)
13. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 48–56. Springer (2019)
14. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* **30**, 5574–5584 (2017)
15. Kohl, S.A., Romera-Paredes, B., Meyer, C., Fauw, J.D., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.A., Rezende, D.J., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 6965–6975 (2018)

16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* **30**, 6402–6413 (2017)
17. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: *International Conference on Learning Representations* (2018)
18. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: *Advances in Neural Information Processing Systems*. pp. 7167–7177 (2018)
19. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: *International Conference on Learning Representations* (2018)
20. Ma, J., Ge, C., Wang, Y., An, X., Gao, J., Yu, Z., Zhang, M., Liu, X., Deng, X., Cao, S., Wei, H., Mei, S., Yang, X., Nie, Z., Li, C., Tian, L., Zhu, Y., Zhu, Q., Dong, G., He, J.: Covid-19 ct lung and infection segmentation dataset (2020). <https://doi.org/10.5281/zenodo.3757476>, <https://doi.org/10.5281/zenodo.3757476>
21. Mehrtash, A., Wells, W.M., Tempny, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging* **39**(12), 3868–3878 (2020)
22. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 12756–12767. Curran Associates, Inc. (2020)
23. Morozov, S., Andreychenko, A., Pavlov, N., Vladzmyrskyy, A., Ledikhova, N., Gomboleviskiy, V., Blokhin, I.A., Gelezhe, P., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465* (2020)
24. Parekh, M., Donuru, A., Balasubramanya, R., Kapur, S.: Review of the chest ct differential diagnosis of ground-glass opacities in the covid era. *Radiology* **297**(3), E289–E302 (2020)
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Louppe, G.: Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12** (01 2012)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 234–241. Springer (2015)
27. Wei, D., Zhou, B., Torrabi, A., Freeman, W.: Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379* (2015)

## Appendix

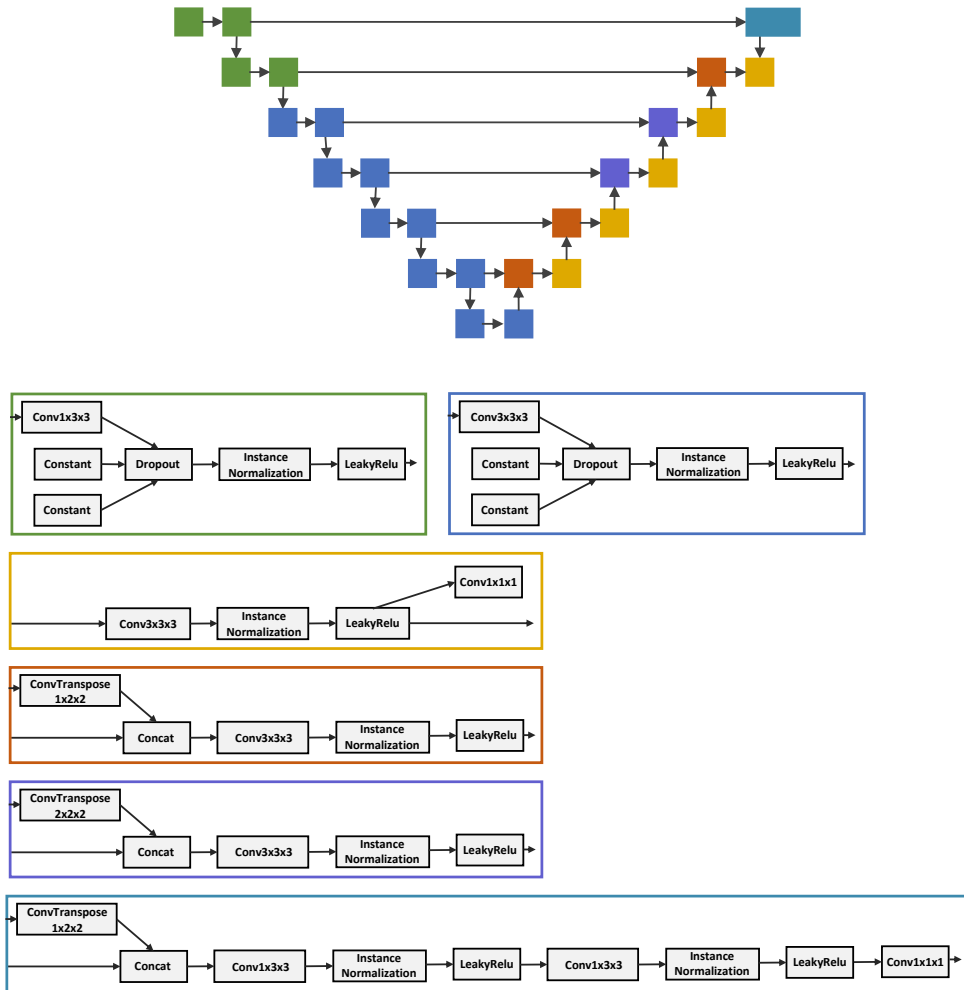
**Table 1.** Summary of characteristics for Covid-19 lung lesion segmentation datasets.

Dataset name	Nr. of subjects	Mean spatial resolution	Mean spacing
Challenge	199	[68.87, 512.0, 512.0]	[4.81, 0.78, 0.78]
Mosmed	50	[40.98, 512.00, 512.00]	[8.00, 0.73, 0.73]
Radiopedia	20	[176.00, 559.55, 571.00]	[1.00, 1.00, 1.00]
In-house	50	[266.64, 819.20, 825.68]	[1.89, 0.60, 0.60]

**Table 2.** Training parameters and specifications. For further details refer to the nnU-Net code at [github.com/MIC-DKFZ/nnUNet](https://github.com/MIC-DKFZ/nnUNet).

Specification	Value
Patch size	[28, 256, 256]
Batch size	250
Padding	None
Loss	Cross-entropy and Dice (smoothing 1e-5), weighted equally
Optimizer	SGD
Initial learning rate	0.01
Weight decay	3e-5
Momentum	0.99
Foreground oversampling	33%
Train-time augmentation	Elastic deformation, scaling, rotation, gamma transformation
Test-time augmentation	None





**Fig. 1.** Model architecture, configured based on properties of the training dataset by the nnU-Net framework.

---

## Contribution and impact

Since the start of the Covid-19 pandemic, many algorithms were proposed for automatic lung lesion segmentation with the hope of alleviating the medical workforce (Liu et al., 2022a). In particular, the *nnU-Net* framework (Isensee et al., 2021) showed promising results in the segmentation of ground glass opacities (GGOs) and consolidations. Yet very few solutions successfully translated to the clinical routine (Hu et al., 2020; Parekh et al., 2020).

Given the insufficient external validation of many methods (Roberts et al., 2021), this may have been the right decision. In our work, we show how a model trained with supposedly heterogeneous data fails to generalize to two other datasets. Low-quality predictions are particularly problematic for lung lesion segmentation, and these have diffuse shapes and can manifest in different sections of the lung, making it difficult to assess their correctness. Network outputs, as we have previously seen, are also not reliable.

In order to identify such *silent failures*, we propose a method that estimates the distribution of training features in a low-dimensional space and, during testing, assesses how far the test activations are in terms of Mahalanobis distance. Though we are not the first to propose this idea (Lee et al., 2018b), our contribution lies in adapting it to semantic segmentation and, specifically, highly parameterized 3D models. To ensure that we can calculate the covariance  $\Sigma$  in an acceptable time frame, we apply pooling operations until the feature size is below  $10^3$ . This allows us to estimate the feature distribution and calculate the Mahalanobis distance purely in the CPU. As we operate in a full-resolution patch-based model, we extract features from each patch, calculate the distance, and add the *distance patch* to an *uncertainty mask* with the size of the original image.

## Discussion and limitations

The method effectively differentiates ID from OOD samples, which allows us to detect silent failures that are the product of domain shift. However, it does not provide an ideal model calibration. In particular, it fails to identify OOD images for which the model *does* produce high-quality segmentation masks.

Another central limitation of the work is the experimental scope, where we focused only on the segmentation of Covid-19 lesions in chest CTs. An additional concern lies in the fact that, as we used two openly available datasets, we could not assess which factors caused the shift in distribution. These could come from a change in the population (likely, as the data came from different global regions including Russia, China, and Germany), acquisition practices (which also varied between data sources), and/or variation in the annotation protocols. We address these concerns in our follow-up paper described in the next section, where we validate our method across a number of scenarios and on MRI data.

### 3.1.3. Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation

We significantly extended our work presented in the previous section for a MICCAI 2021 Special Edition in the *Medical Image Analysis* journal. We titled this new publication *Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation* (González et al., 2022a). The version of the paper contained in this thesis was published on August 24<sup>rd</sup>, 2022. A pre-print with very similar content but minor style changes is available in *arXiv* since August 5<sup>th</sup>, 2022.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation" (González et al. 2022) was published as a full research paper at the "Medical Image Analysis". It constitutes a joint work of Camila González, Karol Gotkowski, Moritz Fuchs, Andreas Bucher, Armin Dadras, Ricarda Fischbach, Isabel Kaltenborn and Anirban Mukhopadhyay.*

*This work was supported by the RACOON network under BMBF, grant number [01KX2021]; and the Bundesministerium für Gesundheit (BMG) with grant [ZMV11-2520DAT03A].*

*As corresponding and leading author, C. González led the overall research design, management and writing process of the paper. M. Fuchs was responsible for the literature review. The research design and choice of methodological framework was done by C. González with supervision from A. Mukhopadhyay. A. Bucher, R. Fischbach and I. Kaltenborn collected the in-house data and reviewed the manuscript from a clinical perspective. A. Dadras contributed to the pre-processing of in-house data, carried out experiments and communicated the results. C. González and K. Gotkowski prepared and pre-processed the openly available data and implemented the code. C. González conducted the experiments and wrote the methodology, results and conclusions. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor of this work, who contributed with continuous feedback during all phases of the paper writing process. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum:	01 / 10 / 2023	01 / 10 / 2023	01 / 10 / 2023	01 / 11 / 2023
Unterschrift:				
	Camila González	Karol Gotkowski	Moritz Fuchs	Andreas Bucher
Datum:	01 / 10 / 2023	01 / 10 / 2023	01 / 10 / 2023	01 / 12 / 2023
Unterschrift:				
	Armin Dadras	Ricarda Fischbach	Isabel Kaltenborn	Anirban Mukhopadhyay



Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: [www.elsevier.com/locate/media](http://www.elsevier.com/locate/media)

## Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation

Camila González<sup>a,\*</sup>, Karol Gotkowski<sup>a</sup>, Moritz Fuchs<sup>a</sup>, Andreas Bucher<sup>b</sup>, Armin Dadras<sup>b</sup>, Ricarda Fischbach<sup>b</sup>, Isabel Jasmin Kaltenborn<sup>b</sup>, Anirban Mukhopadhyay<sup>a</sup>

<sup>a</sup> Darmstadt University of Technology, Karolinenplatz 5, 64289 Darmstadt, Germany

<sup>b</sup> Uniklinik Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

### ARTICLE INFO

MSC:

68T30

68T37

68T45

Keywords:

Out-of-distribution detection

Uncertainty estimation

Distribution shift

### ABSTRACT

Automatic segmentation of ground glass opacities and consolidations in chest computer tomography (CT) scans can potentially ease the burden of radiologists during times of high resource utilisation. However, deep learning models are not trusted in the clinical routine due to *failing silently* on out-of-distribution (OOD) data. We propose a lightweight OOD detection method that leverages the Mahalanobis distance in the feature space and seamlessly integrates into state-of-the-art segmentation pipelines. The simple approach can even augment pre-trained models with clinically relevant uncertainty quantification. We validate our method across four chest CT distribution shifts and two magnetic resonance imaging applications, namely segmentation of the hippocampus and the prostate. Our results show that the proposed method effectively detects far- and near-OOD samples across all explored scenarios.

### 1. Introduction

Automatic segmentation of lung lesions in chest computed tomography (CT) scans could standardise quantification and staging of pulmonary diseases such as Covid-19 and open the way for more effective utilisation of hospital resources. Ground glass opacities (GGOs) and consolidations are characteristic of pulmonary infections onset by the SARS-CoV-2 virus (Parekh et al., 2020). Since the early phases of the pandemic, many institutions have compiled scans from afflicted patients in intensive care, and some initiatives have publicly released cases with ground-truth delineations from expert thorax radiologists (Roth et al., 2021; Jun et al., 2020; Morozov et al., 2020). Deep learning has shown promising results in segmenting these patterns. Particularly the fully-automatic *nnU-Net* (Isensee et al., 2021) secured top spots (Henderson, 2021) (9 out of 10, including the first) in the leaderboard for the *Covid-19 Lung CT Lesion Segmentation Challenge* (Roth et al., 2021).

Unfortunately, models trained with publicly available cohorts may not generalise well to real-world clinical data, thus posing safety issues when deployed without extensive testing and/or quality assurance (QA) protocols. Deep learning models are known to fail for data that diverges from the training distribution (Mehrtash et al., 2020); a phenomenon commonly referred to as *domain shift*. This hinders the deployment of AI solutions during the Covid-19 pandemic (Hu et al., 2020), as most

institutions do not dedicate resources to annotate in-house datasets. There are many potential causes for domain shift, ranging from changes in the acquisition process to naturally shifting patient populations. Some can unknowingly occur within the same institution, rendering even models trained with in-house data unreliable with the passage of time (Srivastava et al., 2021).

This performance deterioration is visualised in Fig. 5 for an *nnU-Net* trained on data from the *COVID-19 Lung CT Lesion Segmentation Challenge* (Roth et al., 2021; An et al., 2020; Clark et al., 2013). Featuring 199 cases, 160 of which were used for training, the data pool is much larger than single institutions realistically collect and annotate, considering how time-intensive the process of lung lesion delineation is. The data is also multi-centre and diverse with regard to patient group and acquisition protocol, yet the model fails to generalise to different distribution shifts. Lung lesions do not manifest in large connected components (see Fig. 12), so it is not trivial for novice radiologists to identify incorrect segmentations.

While we have so far painted a sombre outlook for clinical use of deep learning models, these could still be safely utilised alongside proper quality assurance mechanisms. The problem is that human-performed QA is time-consuming and expensive, ultimately defeating the promise of AI in radiology. On the other hand, automatic methods

\* Corresponding author.

E-mail address: [camila.gonzalez@gris.tu-darmstadt.de](mailto:camila.gonzalez@gris.tu-darmstadt.de) (C. González).

may be an inexpensive and effective first step in identifying low-quality cases. In particular, reliable *out-of-distribution* (OOD) detection can signal when the model is unsuitable for a patient.

Existing methods for OOD detection or uncertainty quantification either (a) observe the network logits, which often *fail silently* exhibiting plausible behaviour mimicking in-distribution (ID) cases even for novel inputs (Hein et al., 2019) or (b) require special training considerations that reduce their usability, such as a self-supervision loss term or outlier detector. In practice, models are used which exhibit the best performance in the target task. Widely-used segmentation frameworks *are not designed with OOD detection in mind*, and so a method is needed that reliably identifies OOD samples post-training while requiring minimal intervention.

We propose to directly estimate the similarity of new samples to the training distribution in a low-dimensional feature space. A large distance signals that the model has not seen specific activation patterns in the past, and therefore outputs produced from such novel features *cannot be trusted*. Our method (Gonzalez et al., 2021), initially presented at MICCAI 2021, is lightweight and requires no changes to the network architecture of the training procedure, allowing it to integrate into complex segmentation pipelines seamlessly. Further, as the distance estimation process follows after training, it can provide clinically-relevant uncertainty scores for pre-trained models.

Building on our previous work, in the present article we provide more context into our methodology, perform an ablation study on selecting feature maps and considerably extend our evaluation. We validate our proposed method across *four* scenarios with a nnU-Net trained on *Challenge* data.

1. For the first setting, we perform inference on the publicly available *Radiopedia* and *Mosmed* datasets. This setting, which we have explored in the past, simulates a *dataset shift* situation where the user does not know exactly which changes are introduced.
2. Secondly, we apply affine transformations and synthetic artefacts to the ID test data in order to simulate, respectively, geometric changes in the subject population and common quality problems in CT acquisition.
3. We also evaluate a *diagnostic shift* scenario on an in-house data cohort with 50 Covid-19 and 50 new non-Covid pneumonia patients.
4. Finally, we carry out a *far-OOD* evaluation where we feed colon and spleen CT examinations from the *Medical Segmentation Decathlon* (MSD) to the model.

In addition, we explore two additional segmentation tasks to assess the transferability of our method to other settings, namely hippocampus and prostate segmentation from, respectively, T1- and T2-weighted Magnetic Resonance Images (MRIs). We also perform experiments on a HighResNet (Li et al., 2017) architecture, which does not follow the classic encoder–decoder structure.

Our results show that our proposed distance-based method reliably detects out-of-distribution samples that other approaches fail to identify across a wide array of use cases.

## 2. Related work

Several strategies have shown acceptable OOD detection performance in classification tasks. *Output-based* methods assess the confidence of the logits by estimating their distance from a one-hot encoding. Hendrycks and Gimpel (2017) propose using the maximum softmax output as an OOD detection baseline. Guo et al. (2017) find that replacing the regular softmax function with a *temperature-scaled* variant produces truer estimates, and Liang et al. (2018) complement this approach by adding perturbations to the network inputs. Similarly, Liu et al. (2020b) use *Energy Scoring* to detect OOD samples in a post-hoc fashion. Given access to explicit OOD samples, training

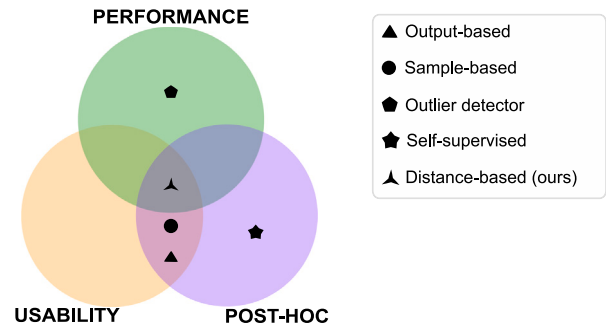


Fig. 1. Desirable properties for OOD detection and corresponding paradigms. A method should ideally (1) be widely applicable (2) work on a *post-hoc* basis even if OOD detection was not a goal during training and (3) reliably detect OOD samples.

Table 1

Comparison between Output- (O), Sample- (S) and Distance-based (D) methods. We compare important factors for applicability: parameters, number of modifications (0–3) and additional inference time from high [–] to none [++].

Method	Type	Parameters	Mod. level	Inf. time
Max. Softmax	O	t	0	++
Temp. Scaling	O	t, T	1	++
KL	O	t, $p(\theta)$	2	+
Energy Scoring	O	t, T	1	++
MC Dropout	S	t, p	3	–
TTA	S	t, $I_{Aug}$	2	– –
<b>Ours</b>	D	t, $\mu, \sigma$	2	+

with an energy-based loss can further improve OOD detection. Other methods (Hendrycks et al., 2019; Lee et al., 2018a) instead look at the KL divergence of softmaxed outputs from the uniform distribution.

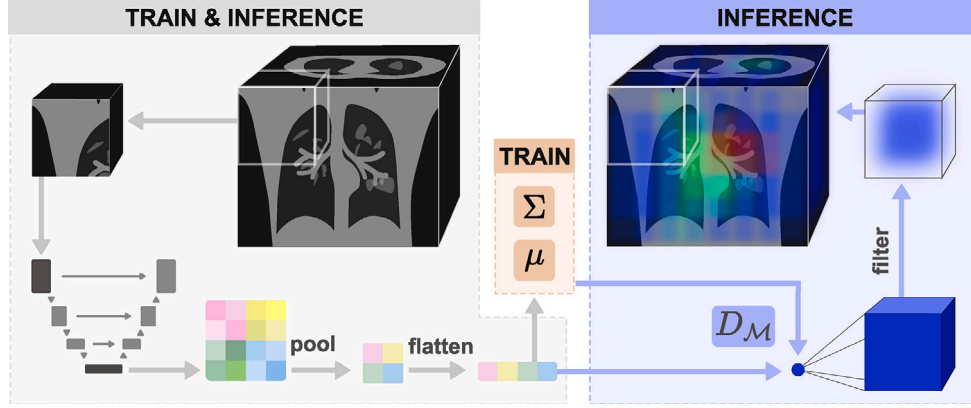
*Sample-based* Bayesian-inspired techniques (Blundell et al., 2015) consider the divergence between several outputs produced under different conditions as the uncertainty. Commonly-used methods are Monte Carlo Dropout (MC Dropout) (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017). The latter usually performs better but requires several models to be trained, whereas MC Dropout can assess uncertainty for any model trained with Dropout layers. Ashukha et al. (2019) show that Test-Time Augmentation (TTA) can significantly improve both singular models and ensembles. Sample-based methods have shown promising results in the field of medical image segmentation (Jungo et al., 2020; Jungo and Reyes, 2019; Mehrtash et al., 2020).

Other approaches use OOD data to explicitly train an *outlier detector* (Bevandić et al., 2019; Hendrycks et al., 2018; Lee et al., 2018a). However, as they require OOD detection to be a primary goal throughout the training process, they cannot be applied post-hoc to pre-trained models.

Methods that modify or make certain assumptions on the architecture or training procedure have shown good performance (Kohl et al., 2018; Monteiro et al., 2020a,b; Fuchs et al., 2021). For instance, *self-supervision* losses provide valuable assessments for novelty (Pidhorskyi et al., 2018; Golan and El-Yaniv, 2018; Hendrycks et al., 2019; Gonzalez and Mukhopadhyay, 2021). However, their applicability to widely-used segmentation frameworks – which do not typically use self-supervision – is limited.

In Fig. 1, we illustrate how existing paradigms perform in terms of different desiderata. We are interested in approaches that can be directly used with any model, and so we restrict our analysis to the methods outlined in Table 1.

Unlike previous work, our method observes model activations at the end of the encoder. We project these to a lower-dimensional feature space and estimate a multi-variate Gaussian with the training data. During inference, we detect samples with a high *Mahalanobis* distance to this distribution, which is suitable for quantifying differences in the latent space (Lee et al., 2018b; Çallı et al., 2019).



**Fig. 2.** Proposed method for OOD detection on a full-resolution nnU-Net model. The input image first goes through a series of pre-processing steps and is divided into patches. For each patch, we take the feature maps generated at the end of the encoder during the forward pass. We then project these into a lower-dimensional, flattened subspace. During the training phase, we estimate a Gaussian distribution from the feature space by calculating  $\mu$  and  $\Sigma$ . At inference time, we calculate the Mahalanobis distance to the training distribution and project the resulting point value into the dimensions of the original patch. Finally, a filtering operation is performed to weigh voxels at the centre more heavily, and the result is aggregated into a volume with the same dimensionality as the input image.

### 3. Material and methods

Our proposed method, visualised in Fig. 2, assesses the uncertainty as the distance of new samples to the training distribution in the feature space. First, we extract feature maps from the trained model and project these to a low-dimensional space to ensure a computationally inexpensive calculation. We then estimate a multi-variate Gaussian distribution from ID train samples. At test time, we repeat the feature-extraction process and calculate the Mahalanobis distance.

We first briefly introduce the patch-based nnU-Net architecture in Section 3.1 and outline how our method links to it. In Section 3.2 we describe our proposed method for OOD detection, which follows a *three-step process*: (1) estimation of a Gaussian distribution from training features (2) extraction of uncertainty masks for test images and finally (3) calculation of subject-level uncertainty scores.

#### 3.1. Patch-based nnU-Net

The nnU-Net is a standardised framework for medical image segmentation (Isensee et al., 2021) that has reported state-of-the-art results across several benchmarks and challenges (Henderson, 2021). Without deviating from the traditional U-Net structure (Ronneberger et al., 2015), it automatically chooses the best architecture and learning configuration for the training data. The framework also performs pre- and post-processing steps during both training and inference, such as adapting voxel spacing and normalising the intensities.

We use the patch-based full-resolution variant, which is recommended for most applications (Isensee et al., 2021). After performing all necessary preprocessing operations, input image  $x$  is divided into patches following a sliding window approach with an overlap of 50%. This results in  $N$  patches  $\{x_i\}_{i=1}^N$ . A forward pass is made for each patch, at which point we extract feature maps for our method. Predictions for each patch are multiplied by a filtering operation that weights centre-voxels more heavily. Finally, weighted predictions are aggregated into an output mask with dimensionality of the original image.

We also experiment with a 3D HighResNet model (Li et al., 2017), which we integrate into the nnU-Net framework and thus follow the same steps for image preparation and combination of the outputs into a coherent prediction.

#### 3.2. Distance-based OOD detection

We are interested in capturing *epistemic uncertainty*, which arises from a lack of knowledge about the data-generating process. While

most uncertainty estimation methods quantify this uncertainty for prediction *boundaries*, we want to do so for whole *regions*, which is challenging for OOD data (Kendall and Gal, 2017).

One way to directly assess epistemic uncertainty is to calculate the distance between training and testing activations. As a model is unlikely to produce reasonable outputs for features far from any seen during training, this is a reliable signal for bad model performance (Lee et al., 2018b).

Model activations have covariance, and they do not necessarily resemble the mode for high-dimensional spaces (Wei et al., 2015), so the Euclidean distance is not appropriate for identifying unusual activation patterns. Instead, inspired by the work of Lee et al. (2018b), we make use of the *Mahalanobis distance*  $D_M$ , which rescales samples into a space without covariance. Fig. 3 illustrates how the Mahalanobis distance better captures the behaviour of in-distribution data and correctly identifies samples outside the unit circle as OOD.

The following sections describe how we leverage the Mahalanobis distance in our approach. Note that only one forward pass is necessary for each patch, keeping the computational overhead at a minimum.

##### 3.2.1. Estimation of the training distribution

We start by estimating a multivariate Gaussian distribution  $\mathcal{N}(\mu, \Sigma)$  over training features. For all training patches  $\{x_i\}_{i=1}^N$ , features  $\mathcal{F}(x_i) = z_i$  are extracted from the encoder  $\mathcal{F}$ .

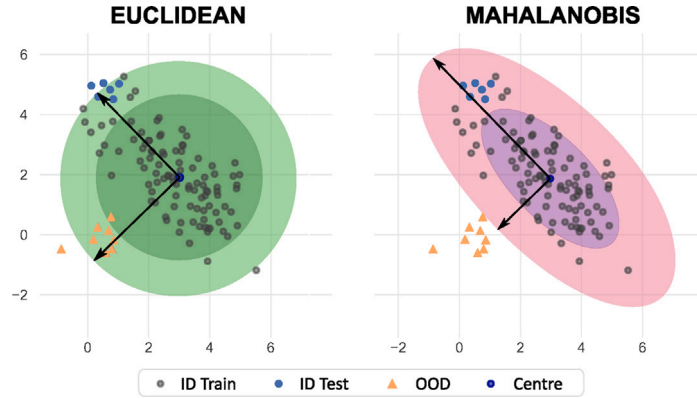
For modern segmentation networks, the dimensionality of the extracted features  $z_i$  is too large to calculate the covariance  $\Sigma$  in an acceptable time frame. We thus project the latent space into a lower subspace by applying average Pooling operations with a kernel size of (2, 2, 2) and stride (2, 2, 2) until the dimensionality falls below  $1e4$  elements. Finally, we flatten this subspace and estimate the empirical mean  $\mu$  and covariance  $\Sigma$ .

$$\mu = \frac{1}{N} \sum_{i=1}^N \hat{z}_i, \quad \Sigma = \frac{1}{N} \sum_{i=1}^N (\hat{z}_i - \mu)(\hat{z}_i - \mu)^T \quad (1)$$

In Table 2 we demonstrate that for a dimensionality of  $1e4$  elements we can estimate the covariance in a maximum of a few minutes (rows 3 and 4) with the *Scikit Learn* on an AMD Ryzen 9 3900X CPU, whereas for higher dimensions the times increase abruptly (row 5).

##### 3.2.2. Extraction of uncertainty masks

During inference, we estimate an uncertainty mask for a subject following the process illustrated in Fig. 2 (right). First, we perform the same preprocessing steps as during training and divide the image into patches. Next, we extract feature maps for each patch  $x_i$  and project them onto  $\hat{z}_i$  as done during training. We then calculate the



**Fig. 3.** Comparison between Euclidean and Mahalanobis distances in a two-dimensional space. Left: Euclidean distance fails to detect that OOD samples (orange triangles) strongly deviate from the expected behaviour of training samples (grey circles). Right: Mahalanobis distance adequately detects OOD samples, assigning them a distance outside the unit circle whilst properly admitting ID test samples (blue circles).

**Table 2**

Times in seconds required for estimating the covariance  $\Sigma$  (column 3) and calculating the Mahalanobis distance  $D_M$  to one sample (column 4).

Nr. samples	Dimensionality	$\Sigma$ time (s)	$D_M$ time (s)
1e3	1e3	0.260	0.001
1e6	1e3	8.480	0.001
1e3	1e4	69.11	0.050
1e4	1e4	81.80	0.051
1e3	2e4	6555.13	0.194

Mahalanobis distance (Eq. (2)) to the Gaussian distribution estimated in the previous step.

$$D_M(\hat{z}_i; \mu, \Sigma) = (\hat{z}_i - \mu)^T \Sigma^{-1} (\hat{z}_i - \mu) \quad (2)$$

Each distance is a point estimate for the corresponding patch. We replicate this value to the size of the patch and combine the distances for all patches in the same manner as the segmentation pipeline combines patch outputs into a coherent prediction.

Following the example of the patch-based nnU-Net, we start by initialising a zero-filled tensor with the dimensionality of the original image. We then apply a filtering operation to each patch to weigh voxels at the centre more heavily and add them to the image-level mask.

### 3.2.3. Subject-level uncertainty

The previous step produces an uncertainty mask with the dimensionality of the input CT scan. In order to effectively identify highly uncertain images, we average over all voxels to obtain one value  $U$ , and normalise uncertainties between the minimum and doubled maximum uncertainties for ID train data to ensure  $U \in [0, 1]$ .

## 4. Experimental setup

We start by describing the data used in our experiments in Section 4.1. Afterwards, we state relevant details on our models (Section 4.2). We then introduce all baselines (Section 4.3) and define our evaluation metrics (Section 4.4).

### 4.1. Data

We train our first model with data from the *COVID-19 Lung CT Lesion Segmentation Challenge* (Roth et al., 2021; An et al., 2020; Clark et al., 2013), which we refer to as *Challenge* or in-distribution (ID). The dataset contains chest CT scans for patients with a confirmed SARS-CoV-2 infection from various centres and countries. The data is also heterogeneous in terms of age, gender, and disease severity of

**Table 3**

Characteristics of the Covid-19 lung lesion segmentation datasets.

Dataset name	Nr. cases	Mean image size	Mean spacing
Challenge	199	[512, 512, 69]	[0.8, 0.8, 4.8]
Mosmed	50	[512, 512, 41]	[0.7, 0.7, 8.0]
Radiopedia	20	[560, 571, 176]	[1.0, 1.0, 1.0]

**Table 4**

Parameters used to randomly generate artefacts and affine transformations with the *TorchIO* library. For each type of shift, three transformed datasets are generated with increasingly stronger transformations.

Shift	Operation	Weak	Medium	Strong
Artefact	Ghost intensity	(0, 0.2)	(0, 0.4)	(0, 0.7)
	Spike intensity	(0, 0.2)	(0, 0.5)	(0, 0.7)
	Blur STD	(0, 0.3)	(0, 0.3)	(0, 0.3)
	Noise STD	(0, 15)	(0, 30)	(0, 30)
Affine	Scales	(0.9, 1.4)	(0.7, 1.8)	(0.6, 2)
	Rotation degrees	5	8	9
	Translation range	(-15, 15)	(-20, 20)	(-20, 20)
	Isotropic	True	True	False

the patients. We use the 199 cases that are made available for the challenge, which we divide into 160 training and 39 testing cases with the nnU-Net random splitting function.

We include results for four types of out-of-distribution samples: (1) **dataset shift**, where we evaluate the model on two other datasets with differences in the acquisition and population patterns (2) **transformation shift** where we apply artificial transformations to our ID data, (3) **diagnostic shift**, where we compare Covid-19 to non-Covid pneumonia patients, and (4) **far-OOD**, where we use the *Spleen* and *Colon* tasks of the Medical Segmentation Decathlon (MSD) (Simpson et al., 2019; Antonelli et al., 2022).

In addition, we perform a study on hippocampus and prostate segmentation from MR images. We train each nnU-Net model with the corresponding task of the MSD and use two and three OOD datasets for hippocampus and prostate, respectively.

#### 4.1.1. Dataset shift

We use two publicly available datasets: *Mosmed* (Morozov et al., 2020) contains fifty cases and the *Radiopedia* dataset (Jun et al., 2020), a further twenty. Both encompass patients with and without confirmed infections. Table 3 provides a summary of data characteristics.

#### 4.1.2. Transformation shift

We transform the 39 in-distribution test cases with multiple operations from the *TorchIO* (Pérez-García et al., 2021) library.

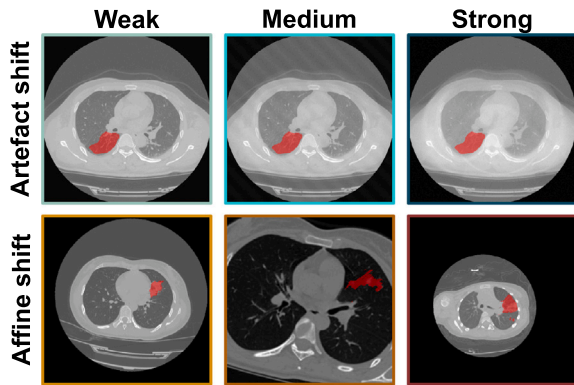


Fig. 4. Top row: Exemplary CT slice with overlaid segmentation mask in red after being transformed to contain artefacts in three magnitudes. Bottom row: Three exemplary CT slices with overlaid segmentation masks after applying affine transformations in three magnitudes. The border colours map each example to their corresponding datasets in Fig. 5.

The *artefact* transformations include ghosting, k-space spikes, Gaussian blurring, and Gaussian noise. *Affine* transformations include scaling, rotation, and translation. All affine operations can be either isotropic or anisotropic. We deploy the same transformation parameters for the sagittal, coronal, and axial dimensions for the isotropic case. For the anisotropic case, these parameters change for every dimension, causing a stronger shift. For both groups of transformations, we generate three sets (*weak*, *medium*, and *strong*), each with increasingly stronger augmentation parameters. The parameters used are reported in Table 4. Examples of the performed transformations are visualised in Fig. 4.

#### 4.1.3. Diagnostic shift

We utilise an in-house dataset of one hundred cases. Fifty patients have pulmonary infection of Covid-19 confirmed by RT PCR test and visible pulmonary Covid-19 lesions in all cases (3/2020 to 12/2020). The remaining fifty cases were composed of various Covid-mimics, manifesting similar pulmonary lesions but acquired prior to the Covid outbreak or tested negative for Covid-19 by RT PCR (3/2017 to 2/2020). Cases were collected and annotated in the RACOON project (Roefo, 2022). Covid-mimics included are viral non-Covid pneumonia, bacterial pneumonia, fungal pneumonia, tuberculosis, chronic obstructive pulmonary disease, cystic fibrosis, interstitial pulmonary fibrosis, acute interstitial pneumonia, cryptogenic organising pneumonia, medication associated pulmonary toxicity, radiogenic pulmonary fibrosis, acute lung embolism, chronic lung embolism, pleural pathologies, pulmonary vasculitis, bronchial carcinoma, pulmonary metastasis, as well as a control case without any lung pathologies.

A clinical radiologist with 8 years of experience in reading chest CT reviewed all scans and found them to be of good enough quality for accurate visual diagnosis. Manual annotations of the entire image stack were performed slice-by-slice by two independent readers trained in the delineation of GGOs and pulmonary consolidations. Central vascular structures and central bronchial structures were excluded from all annotations. Care was taken to differentiate between artefacts and GGO. Consolidations were defined as visible in a soft tissue window and at least 5 mm in size. An expert radiologist reader reviewed all delineations. In Table 5 we report some details on the demographic distribution.

#### 4.1.4. MRI tasks

For hippocampus we consider three T1-weighted datasets: the MSD task, which we denote *MSD H*, and contains healthy and schizophrenia patients, the *Dryad* (Kulaga-Yoskovitz et al., 2015) dataset with fifty

Table 5

In-house data cohort with 50 Covid-19 and 50 non-Covid cases. We report the age (median Q1/Q3), gender (f/m), voltage (median kV), and tube current-time product (mAs).

	Age	Gender	Voltage	mAs
Covid-19	57.17 [49/67]	16%	100	121.21 ± 55.91
Non-Covid	60.24 [47/73]	42%	120	114.77 ± 82.56

Table 6

Characteristics of the MR hippocampus (top) and prostate (bottom) segmentation datasets. Models were trained with the respective tasks of the *Medical Segmentation Decathlon*.

Dataset name	Nr. cases	Mean image size	Mean spacing
MSD H	260	[50, 35, 36]	[1.0, 1.0, 1.0]
Dryad	50	[64, 64, 48]	[1.0, 1.0, 1.0]
HarP	270	[64, 64, 48]	[1.0, 1.0, 1.0]
MSD P	32	[316, 316, 19]	[1.0, 1.0, 1.0]
ISBI	30	[384, 384, 19]	[0.5, 0.5, 3.7]
UCL	13	[384, 384, 24]	[0.5, 0.5, 3.3]
I2CVB	19	[384, 384, 64]	[0.5, 0.4, 1.3]

healthy subjects and the *Harmonised Hippocampal Protocol* data (Boccardi et al., 2015) (*HarP*) with senior subjects, some of which have Alzheimer's.

For the segmentation of the prostate in T2-weighted MRIs we use a corpus of four datasets including the MSD data (*MSD P*) and three OOD sets: the cases provided in the *NCI-ISBI 2013 Challenge* (Bloch et al., 2015) (*ISBI*) and the *I2CVB* (Lemaître et al., 2015) and *UCL* (Litjens et al., 2014) datasets as made available by Liu et al. (2020a). To align label characteristics, we unify the labels of *head* and *body* for the hippocampus and of *central gland* and *peripheral area* for the prostate. A summary of the relevant dataset characteristics can be found in Table 6.

## 4.2. Models

We train three patch-based nnU-Nets (Isensee et al., 2021) and one HighResNet (Li et al., 2017) on a *Tesla T4* GPU. Our configurations have patch sizes of [256, 256, 28], [56, 40, 40] and [320, 320, 20] for the *Challenge*, *MSD H* and *MSD P* tasks, respectively. In all cases, adjacent patches overlap by 50%, and we train with a loss of Dice (smoothing 1e-5) and Binary Cross-entropy weighted equally until after convergence. Training begins with a learning rate of 0.01 and a weight decay of 3e-5. No test-time augmentation was applied to extract predictions, as this signifies a speed-up of 8 times for 3D data.

## 4.3. Baselines

We compare our approach to output- and sample-based techniques that assess uncertainty information by performing inference on a trained model. *Max. Softmax* consists of taking the maximum softmax output (Hendrycks and Gimpel, 2017). *Temp. Scaling* performs temperature scaling on the outputs before applying the softmax operation (Guo et al., 2017). *KL from Uniform* computes the KL divergence from a uniform distribution (Hendrycks et al., 2019). Note that all three methods output a *confidence* score (higher is more certain), which we invert to obtain an *uncertainty* estimate (lower is more certain). *Energy Scoring* (Liu et al., 2020b) assesses uncertainty as the logarithmic sum of the softmax denominator.

*MC Dropout* (Gal and Ghahramani, 2016) consists of doing several forward passes whilst activating the Dropout layers that would usually be dormant during inference. We perform 10 forward passes. Test-Time Augmentation (TTA) follows a similar strategy by augmenting images during testing (Wang et al., 2019). We use image-flip as augmentation and generate eight predictions by flipping the input image once clockwise and counter-clockwise for every axis. We report the standard deviation between outputs as an uncertainty score for both methods.



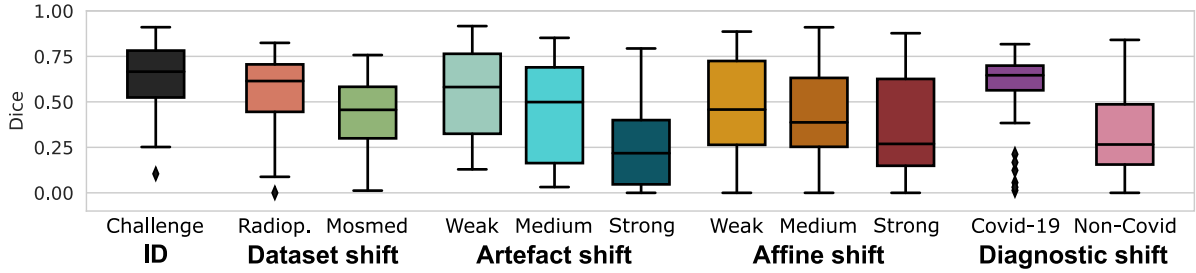


Fig. 5. Performance deterioration of a model trained with ID (*Challenge*) data and tested on (1) *Radiopedia* and *Mosmed*; *Challenge* test cases after applying (2) artefact and (3) affine transformations with different levels of intensity; and (4) in-house Covid-19 and non-Covid pneumonia patients.

For all baselines and our proposed method we calculate a subject-level metric by averaging voxel values, and normalise the uncertainty range between the minimum and doubled maximum uncertainty represented in ID train data. For *Energy Scoring* and *Temp. Scaling*, we always report the result with lowest ESCE from among three different temperature settings  $T \in \{1, 10, 100\}$ .

#### 4.4. Metrics

For OOD detection, we calculate the 95% *true positive rate* (TPR) boundary on ID data, i.e. the boundary that covers at least 95% of train samples. Samples with uncertainties greater than this boundary are predicted to be OOD. We report the *false positive rate*, defined as

$$FPR = \frac{FP}{FP + TN}, \quad (3)$$

where a *false positive* (FP) is an OOD sample incorrectly deemed to be in-distribution, the *Detection Error*

$$Error = \frac{1}{2}(1 - TPR) + \frac{1}{2}FPR \quad (4)$$

and the *area under the receiving operating curve* (AUC), calculated with the *Scikit Learn* library (Pedregosa et al., 2012).

While the detection of OOD samples is a first step in assessing the suitability of a model for a new image, an ideal uncertainty metric would inversely correlate with model performance. For this, we calculate the *Expected Segmentation Calibration Error* (ESCE). Inspired by Guo et al. (2017), we divide the  $n$  test scans into  $M = 10$  interval bins  $B_m$ . For each bin, the absolute difference is calculated between average Dice ( $Dice(B_m)$ ) and inverse average uncertainty ( $1 - U(B_m)$ ) for samples in the bin. A weighted average is reported that weights the score for each bin by the number of samples in it (Eq. (5)).

$$ESCE = \sum_{m=1}^M \frac{|B_m|}{n} |Dice(B_m) - (1 - U(B_m))| \quad (5)$$

## 5. Results

We first analyse the *dataset shift* scenario, where a model trained on the *Challenge* dataset is tested on publicly available *Radiopedia* and *Mosmed* cases (Section 5.1). Afterwards, we evaluate how robust the model is against the presence of artefacts and affine transformations of different magnitudes and explore to what extent these are correctly detected (Section 5.2). As a third setting, we apply our method to an in-house data cohort with both Covid-19 and non-Covid patients in Section 5.3.

In Section 5.4, we perform a *far-OOO* study where we examine whether our method detects samples very far from the raining distribution. We then carry out an ablation study where we measure the use of different network layers for feature extraction (Section 5.5) and repeat the *dataset shift* experiments on a HighResNet model (Section 5.6). In all these experiments, we explore whether our method can distinguish between ID cases – test subjects from the *Challenge* data – and

Table 7

Dataset shift results. Ability of assessing segmentation quality as Estimated Segmentation Calibration Error (ESCE) and identifying samples from *Radiopedia* and *Mosmed* as OOD in terms of Detection Error (Error), False Positive Rate (FPR) and Area Under the ROC (AUC).

Method	ESCE ↓	Error ↓	FPR ↓	AUC ↑
Max. Softmax	.39	.43	.84	.61
MC Dropout	.28	.41	.79	.75
KL	.38	.44	.83	.69
TTA	.36	.41	.77	.74
Temp. Scaling	.02	.47	.89	.42
Energy Scoring	.46	.51	.90	.31
<b>Ours</b>	.15	.09	.04	.96

OOD images. We qualitatively look into exemplary predictions and corresponding uncertainty scores in Section 5.7.

Finally, in Section 5.8, we evaluate the transferability of our method to MR data, where we look at hippocampus and prostate segmentation tasks.

### 5.1. Dataset shift

In Table 7, we report the performance of our proposed method and six other approaches in identifying the OOD samples, i.e. samples from the *Mosmed* or *Radiopedia* datasets for which the model produces unreliable predictions (see Fig. 5). Following previous research in OOD detection (Liang et al., 2018), we find the uncertainty boundary that covers 95% of in-distribution train samples and deem cases with uncertainties beyond the ID 95th percentile threshold as OOD. Our distance-based method is the only approach that successfully flags cases far from the training distribution, as shown by a low detection error and FPR and an AUC close to one.

We plot the Dice score against normalised uncertainty for the three best-performing methods in Fig. 6. The vertical line marks the 95% TPR boundary. We consider predictions with a Dice score lower than 0.6 to be of *low quality* as they diverge significantly from the ground truth (Valindria et al., 2017) and, for the task of Covid-19 lesion segmentation, provide a misleading assessment of the spread of the infection.

The lower left (red) quadrant is critical for the safe use of segmentation models, as it houses *silent failures* for which *low-quality predictions* are made but which are not identified as such. Only our method assigns sufficiently large uncertainty estimates to poorly segmented OOD samples, excluding them from this section. Nevertheless, the upper right (yellow) quadrant shows that our method is too conservative in estimating uncertainties, not identifying samples for which the model produces good segmentations. This overly cautious behaviour potentially leads to an under-utilisation of the model for cases that are technically OOD but have very apparent lesions which are easy to segment; though any amount of safe utilisation is advantageous. Another limitation of the proposed method is that it fails to identify ID samples that the model segments incorrectly due to the lesions being too small or different from

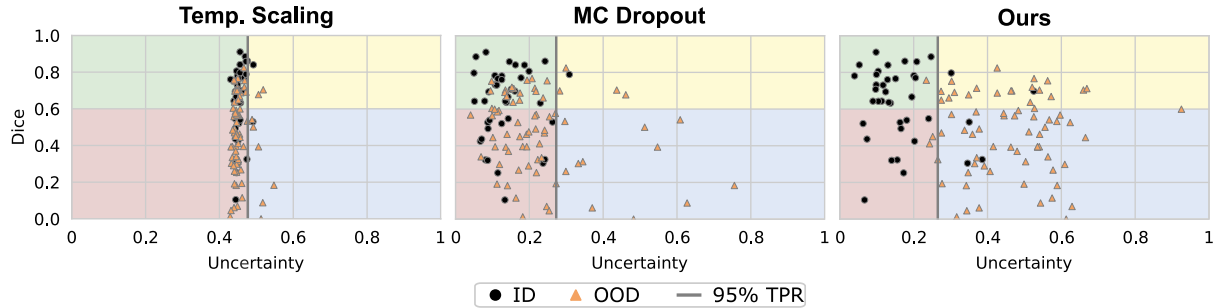


Fig. 6. Dice coefficient against normalised uncertainty for test ID (black circles) and OOD (orange triangles) scans. The ID samples are from the *Challenge* dataset, and the OOD ones from *Mosmed* or *Radiopedia*. The grey vertical line marks the 95% TPR for ID train data. Samples to the right are predicted to be OOD. Clinically relevant is the lower left (red) quadrant that houses silent failures, i.e. predictions with a Dice < 0.6 and low uncertainty scores.

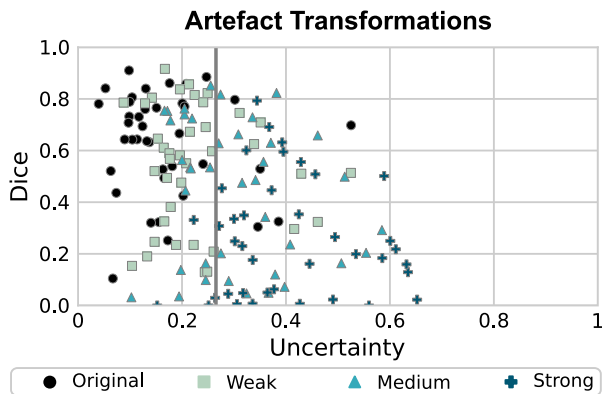


Fig. 7. Dice coefficient against normalised uncertainty. Black circles are the test ID (unmodified *Challenge*) images, and the remaining markers stand for the same *Challenge* images after applying transformations to simulate common artefacts.

those seen in the training data, highlighting the fact that OOD detection is only part of a thorough QA process.

Regarding the estimation of segmentation quality, *Temp. Scaling* reaches the lowest ESCE (first column in Table 7), but a closer inspection of Fig. 6 (left) displays that this is due to most uncertainties clustering on the fifth bin. An ideal segmentation calibration would house all samples in the upper left (green) and lower right (blue) quadrants.

### 5.2. Artefact and affine shifts

The *dataset shift* scenario observed in the previous section depicts a realistic setting whether there are several potential degrees of variation between the training data and cases encountered during deployment. However, it is difficult to assess whether the model performance falls due to (a) changes in the acquisition process, (b) another patient population or simply (c) a different delineation process for ground truth segmentation masks. Subsequently, we cannot confidently assess *why* cases are flagged as OOD. We therefore artificially transform the same ID test cases in two different ways and three levels of magnitude. More than any other explored scenario, these images could be deemed *near-OOD* (Fort et al., 2021). Nevertheless, there is a significant performance deterioration for transformed images, which grows with the magnitude of the perturbation (Fig. 5).

We start by simulating the presence of common image artefacts. In Fig. 7, we visualise the results of our method.

While non-transformed (*original*) cases are correctly assigned low uncertainty scores and most heavily transformed samples are identified as OOD, several samples for which bad segmentations are produced are not identified. Most of these are only weakly transformed (mint-coloured squares). On the other hand, many weakly transformed cases

Table 8

Transformation shift results. Segmentation calibration (as ESCE) and OOD detection scores between original *Challenge* images and cases modified with synthetic artefacts and affine transformations, respectively.

Method	ESCE ↓	Error ↓	FPR ↓	AUC ↑
Max. Softmax	.46/.44	.48/.46	.94/.89	.55/.56
MC Dropout	.44/.44	.51/.51	1.0/.99	.22/.23
KL	.46/.44	.48/.46	.91/.86	.58/.57
TTA	.43/.41	.46/.38	.87/.72	.63/.61
Temp. Scaling	<b>.05/.04</b>	.51/.35	.95/.62	.50/.76
Energy Scoring	.52/.51	.53/.33	.92/.53	.49/.76
<b>Ours</b>	.26/.21	<b>.29/.18</b>	<b>.45/.24</b>	<b>.83/.89</b>

for which good segmentations are produced are correctly assigned low uncertainties despite not being ID. Most heavily transformed images (turquoise crosses) are correctly deemed too far from the training distribution to have reliable predictions.

A similar situation occurs when we apply affine transformations to simulate geometric changes (Fig. 9). These could arise from shifting population patterns, scans being acquired for different ranges, or using other acquisition parameters. Our method deems many weakly transformed cases (yellow squares) to be ID. This is positive as good segmentations are available for most cases. However, a few failure cases are not adequately identified.

Table 8 compares several approaches in terms of OOD detection and segmentation quality assessment. While our method displays an acceptable calibration error and the best OOD detection performance, this *near-OOD* problem proves more difficult than *dataset shift*. It particularly seems to be very difficult to reliably detect image artefacts.

We further visualise the uncertainty ranges assigned to each shift and magnitude in Fig. 8. As expected, the uncertainty increases with the degree of transformation for artefact shifts. For affine shifts, *medium* changes result in similar uncertainties to *strong* ones. This is likely due to the selected transformation sequences being too similar (see Table 4), which results in a similar performance for *medium* and *strong* artefacts (Fig. 5).

In general, we can conclude that the uncertainty correlates positively with the degree of deformation and inversely with model performance. Affine transformations also have a more pronounced effect on the uncertainties (Fig. 8). This possibly stems from the training data containing similar patterns to those introduced by the weaker artefact transformations.

### 5.3. Diagnostic shift

We have not yet analysed how the segmentation model performs across disease patterns. To explore this, we segment lung lesions in the form of GGOs and consolidations for an in-house cohort of 50 Covid-19 and 50 non-Covid cases. The performance of the model on the non-Covid cases is significantly worse. Table 9 summarises our findings, and we plot our uncertainty assessment in Fig. 10.

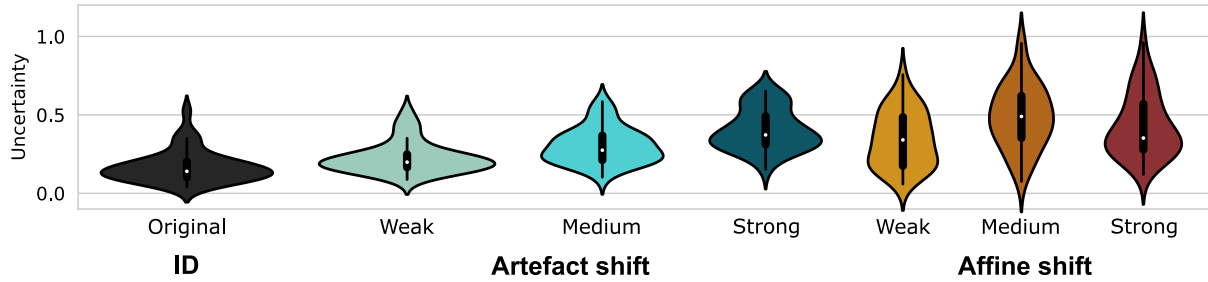


Fig. 8. Distribution of uncertainty scores estimated by our proposed method for the *artefact shift* and *affine shift* scenarios. In general, the uncertainties increase with the intensity of the transformations.

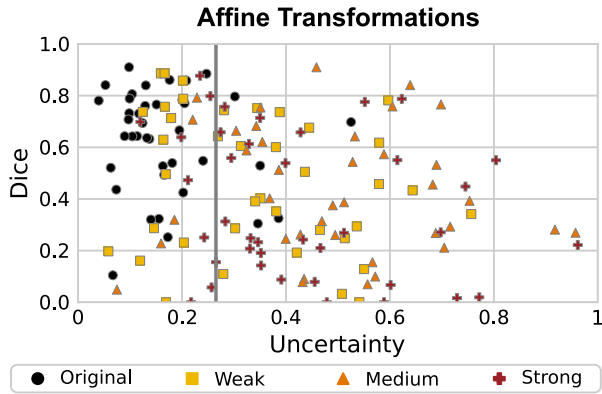


Fig. 9. Dice coefficient against normalised uncertainty. Black circles are the test ID (unmodified *Challenge*) images, and the remaining markers stand for the same *Challenge* images after applying transformations to simulate affine shifts.

Table 9

Diagnostic shift results. Segmentation calibration (as ESCE) and OOD detection scores between test ID *Challenge* images and in-house cases with and without Covid-19, respectively.

Method	ESCE ↓	Error ↓	FPR ↓	AUC ↑
Max. Softmax	.29/.42	.22/.32	.42/.62	.86/.87
MC Dropout	.22/.38	.30/.46	.58/.90	.84/.69
KL	.29/.42	.23/.33	.40/.60	.88/.89
TTA	.25/.32	.19/.17	.32/.28	.89/.95
Temp. Scaling	<b>.07/.05</b>	.34/.54	.62/1.0	.78/.06
Energy Scoring	.38/.54	.49/.56	.86/1.0	.61/.05
<b>Ours</b>	<b>.16/.26</b>	<b>.13/.15</b>	<b>.14/.18</b>	<b>.93/.92</b>

Our method reliably detects cases from our in-house cohort, though it does not distinguish between Covid-19 and non-Covid cases. Though ideally Covid-19 cases for which good predictions are produced should be deemed low-uncertainty, the fact that badly segmented non-Covid cases are flagged as OOD is more relevant for clinical use as unsure good predictions are preferred over confident faulty ones.

5.4. Far-OOD examinations

We have extensively examined *near-OOD* (Fort et al., 2021) cases where a performance deterioration is unexpected. In contrast, *far-OOD* situations occur when an input is erroneously fed into a model, and there is no realistic expectation that a model can produce a sensible prediction.

In Table 10, we examine what happens when we feed CT spleen and colon cancer examinations from the *Medical Segmentation Decathlon* into our model trained to segment pulmonary lesions from chest CTs. Our method distinguishes between ID and far-OOD cases, correctly identifying all colon examinations as OOD (FPR = 0) and showing detection errors of up to 0.1 for both anatomies.

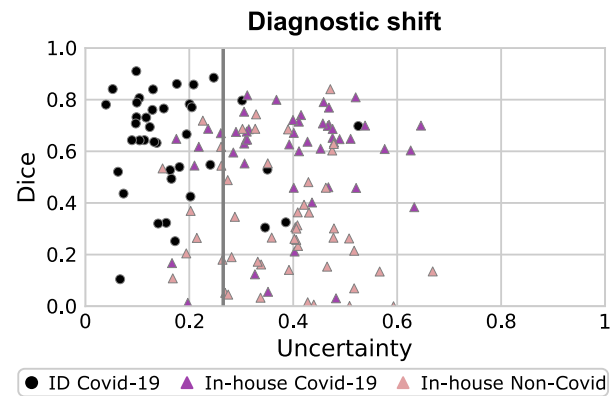


Fig. 10. Dice coefficient against normalised uncertainty for ID test (*Challenge*) data and in-house chest CTs of Covid-19-positive (purple triangles) and non-Covid (pink triangles) patients.

Table 10

Far-OOD results. Segmentation calibration (as ESCE) and OOD detection scores between test ID *Challenge* images and CT scans for spleen and colon examinations, respectively.

Method	ESCE ↓	Error ↓	FPR ↓	AUC ↑
Max. Softmax	.58/.71	.44/.42	.85/.81	.89/.89
MC Dropout	.50/.64	.37/.36	.68/.66	.88/.87
KL	.59/.72	.44/.42	.85/.81	.88/.88
TTA	.48/.58	.18/.22	.29/.37	.95/.95
Temp. Scaling	.62/.71	.48/.42	.93/.81	.79/.89
Energy Scoring	<b>.31/.16</b>	.49/.51	.93/1.0	.50/.50
<b>Ours</b>	<b>.34/.41</b>	<b>.10/.06</b>	<b>.07/.00</b>	<b>.96/.98</b>

5.5. Ablation study

We evaluate which features are most expressive for detecting distribution shifts in Table 11. We compare the use of activations at the middle of the network, more specifically the convolutional (Conv) parameters of the sixth *encoding block* (EB) against those of the first *decoding block* (DB), and features at the beginning (1st EB) and final end (6th DB) of the architecture. In addition, we look into the use of *batch normalisation* (BN) layers, as these normalise layer inputs and therefore contain domain information (Ioffe and Szegedy, 2015). The results show that features at the middle of the network (6th EB Conv, followed by 6th EB BN and 1st DB Conv) are the most suitable for detecting distribution shifts.

5.6. HighResNet model

Not all segmentation models follow an encoder–decoder structure. For instance, the HighResNet (Li et al., 2017) uses dilated convolutions and residual blocks to produce accurate segmentations. That raises the

Table 11

Ablation study on the usability of feature maps. OOD detection and segmentation calibration for our proposed method using different convolutional (Conv) and batch normalisation (BN) at different encoding (EB) and decoding blocks (DB). The results are for the *dataset shift* and *transformed* (including both artefact and affine shifts) scenarios, respectively.

Features	ESCE ↓	Error ↓	FPR ↓	AUC ↑
<b>6th EB Conv</b>	<b>.15/.23</b>	<b>.09/.24</b>	<b>.04/.35</b>	<b>.96/.86</b>
6th EB BN	.18/.23	.11/.25	.09/.37	.95/.85
1st EB Conv	.42/.24	.56/.70	.13/.40	.81/.21
1st EB BN	.52/.45	.50/.50	<b>.00/.00</b>	.51/.51
1st DB Conv	.17/.25	.09/.25	.06/.38	<b>.96/.84</b>
6th DB Conv	.52/.45	.50/.50	<b>.00/.00</b>	.50/.50

Table 12

HighResNet results. Segmentation calibration (as ESCE) and OOD detection scores between test ID *Challenge* images and OOD samples belonging to the *Radiopedia* or *Mosmed* datasets, for a HighResNet model trained on *Challenge*. The bottom part of the table shows three variations of our method with different feature maps: the 7th conv. block, the 6th block with dilated conv., and the 12th (last) block with dilated convolutions.

Method	ESCE ↓	Error ↓	FPR ↓	AUC ↑
Max. Softmax	.35	.48	.94	.57
MC Dropout	.35	.49	.96	.59
KL	.34	.46	.90	.60
TTA	.35	.48	.90	.61
Temp. Scaling	.35	.48	.93	.54
Energy Scoring	.58	.49	.97	.50
7th Conv Block	.41	.47	<b>.00</b>	<b>.94</b>
6th Dil Conv Block	.58	.50	<b>.00</b>	.50
<b>12th Dil Conv Block</b>	<b>.33</b>	<b>.37</b>	<b>.00</b>	.84

questions of whether our proposed approach would be effective on this architecture and which features would be most helpful for detecting distribution shifts. We report these results for the *dataset shift* scenario in Table 12. The upper section summarises the results for all baselines, and the lower part shows the performance of our proposed method for three different feature maps.

The HighResNet architecture is divided into four sections: (1) seven convolutional blocks, (2) six blocks with dilated convolutions using a dilation factor of 2, (3) six dilated convolutional blocks with a factor of 4, and (4) a final convolutional block. Residual connections with identity mapping are also included every two blocks to join features at different levels. We test the use of three feature maps: the last (7th) convolutional block, the last (6th) dilated convolutional block with factor 2, and the last (12th) dilated convolutional block.

The best results are for the variant of our method which uses the last block with dilated convolutions. Though the FPR and AUC are encouraging, the detection error is relatively high, suggesting that the TPR is low as the 95% TPR on ID train data does not cover a significant portion of ID test samples (see Eq. (4)). We plot the performance of the network vs. normalised uncertainties for the best-performing features in Fig. 11. A separation is noticeable between ID (*Challenge*) and OOD (*Radiopedia* and *Mosmed*), but the uncertainty boundary – as hypothesised from the high Detection Error – is too low. This means that OOD samples are correctly detected, yet the model is under-utilised.

### 5.7. Qualitative evaluation

We now take a detailed view of some cases in Fig. 12. The first column shows an in-distribution *Challenge* case with a good prediction. The second and third cases are from *Mosmed* and *Radiopedia*, respectively. While the *Mosmed* prediction is significantly different from the ground truth (incorrectly marking several regions as lesions), a good segmentation is produced for the third case.

We first notice the complexity of assessing whether a segmentation mask for lung lesions is correct. An untrained observer would not be able to detect that the second segmentation is so different from the

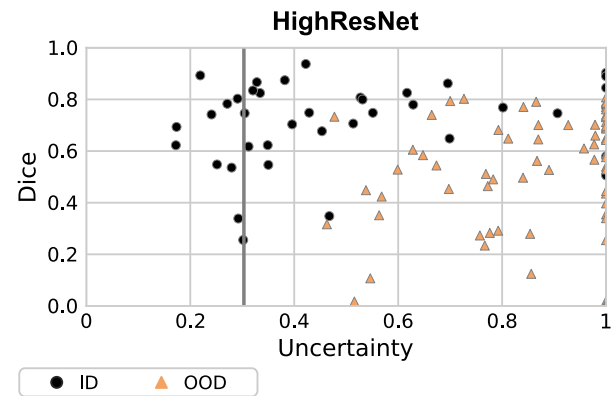


Fig. 11. Dice coefficient against normalised uncertainty for the variant using the 12th Dil. Conv. Block. Black circles are test ID (*Challenge*) images, and orange triangles are OOD cases from *Radiopedia* or *Mosmed*.

ground truth, and even trained radiologists may not directly identify this error, as GGOs can manifest in superior lobes and with multiple connected components (Parekh et al., 2020). Similarly, all methods fail to detect this case except for our distance-based method, which assigns an uncertainty of 0.61.

The prediction for the third case over-segments some lesions, though if we observe the difference between the *Challenge* and *Radiopedia* ground truth masks, we notice that delineations are coarser for the first case (we see in the first image that broad regions around lesions are marked as infected). Therefore, the model learns to mimic this behaviour. Beyond this, the segmentation model correctly detects all lesions and only creates a very small additional component. Here, our method makes an overly cautious uncertainty assessment, assigning this case an uncertainty of .43 which falls beyond the 95% TPR boundary.

### 5.8. Application to MRI data

Magnetic Resonance Imaging (MRI) data is even more susceptible to changes in the acquisition conditions than CTs, as there is no consensus on the calibration of intensity values. This causes the performance of segmentation models trained on MR tasks to deteriorate on OOD data (Zakazov et al., 2021; Kondrateva et al., 2021).

In this section, we evaluate how our proposed method can help detect such distribution shifts on nnU-Net models trained with the *hippocampus* and *prostate* tasks of the MSD. Fig. 13 illustrates that while the initial performance of the models is over 0.8 Dice on in-distribution test data (*MSD H* and *MSD P*), it falls significantly for the OOD datasets.

Table 13 summarises our results on OOD detection, and we visualise the uncertainties of our method in Fig. 14. We immediately see that – for both MR segmentation tasks – detecting OOD cases is much easier than for chest CT. In all cases, the proposed method correctly distinguishes ID from OOD data. This is likely due to the inherent variability across MRI datasets in terms of intensity histogram and fields-of-view. The last row includes a *far-OOD* case where we look to detect *MSD H* cases on the model trained with *MSD P* and vice versa. This also seems to be an easy problem, and our method correctly identifies all OOD cases.

## 6. Discussion

Uncertainty quantification is an unavoidable cornerstone for safely deploying predictive models in real clinics. Our results show that the proposed distance-based approach provides valuable information for detecting images that the model is unprepared to segment.

As distance-based OOD detection can seamlessly augment any segmentation pipeline, there is no reason against performing this quality

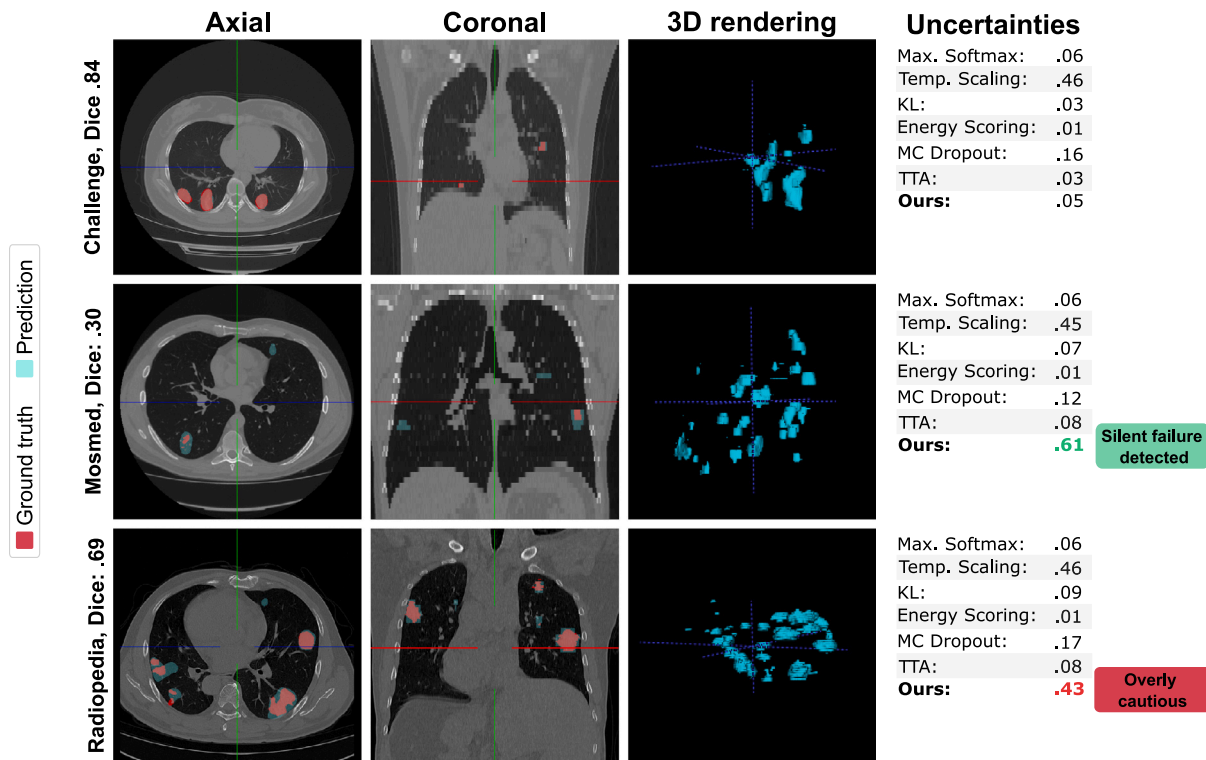


Fig. 12. Axial and coronal slices with overlaid predictions and ground truths and volume renderings of the predictions for three different subjects. First column: a good prediction. Second column: a poor prediction for an OOD case which our method successfully detects. Though there are considerable differences to the ground truth, these errors are not directly noticeable even for trained observers. Third column: a good prediction for an OOD case.

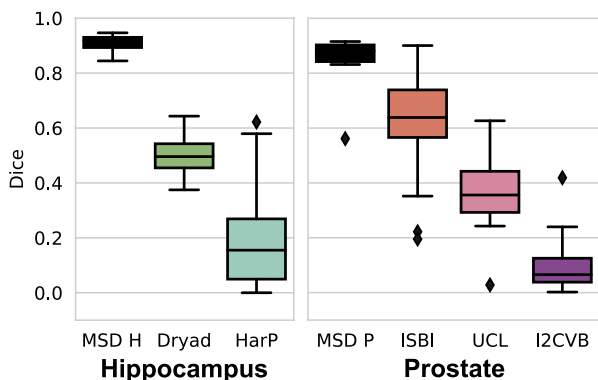


Fig. 13. Performance as Dice score of models trained with *MSD H* (left) and *MSD P* (right) data for hippocampus and prostate segmentation, respectively. Plotted are the ID test (in dark blue) and OOD scores.

check. However, we found in our analysis several areas where there is room for improvement. Almost all our experiments showed that our method is overly cautious in its uncertainty estimation. Specifically, many OOD cases for which the model *did* produce adequate segmentation were deemed highly uncertain. Only for the *artefact shift* scenario were weekly transformed samples segmented.

The *artefact* and *affine shifts* experiments show that – for both explored synthetic scenarios – the produced distances grow linearly with the degree of change and are inversely proportional to segmentation quality. This is ideal behaviour for an uncertainty metric. However, the same does not hold for the *dataset shift* and *diagnostic shift* settings. Particularly for the last scenario, our method assigns similar uncertainties to both Covid-19 and non-Covid cases, even though segmentations are much worse for the last group. Further research should explore which

Table 13

MRI results. Segmentation calibration (as ESCE) and OOD detection scores between test ID and OOD cases for hippocampus and prostate, respectively. The networks were trained with *MSD H* and *MSD P* data, respectively, so these cases are ID. The last row summarises the results for the far-ODD case of detecting *MSD P* cases on the *MSD H* model and vice versa.

Method	ESCE ↓	Error ↓	FPR ↓	AUC ↑
Max. Softmax	.20/.36	.05/.49	.00/.82	1.0/.74
MC Dropout <i>N</i> = 10	.53/.08	.50/.01	1.0/.02	.40/1.0
MC Dropout <i>N</i> = 100	.48/.14	.53/.00	1.0/.00	.12/1.0
KL	.18/.15	.05/.16	.00/.16	1.0/.83
TTA	.20/.40	.09/.25	.00/0.0	1.0/.83
Temp. Scaling	.12/.36	.03/.49	.00/.82	1.0/.74
Energy Scoring	.68/.53	.50/.49	1.0/.98	.50/.12
Ours	.21/.19	.00/.00	.00/.00	1.0/1.0
Ours far-ODD	.08/.01	.00/.00	.00/.00	1.0/1.0

distribution shifts negatively affect model performance, and how these can be distinguished from harmless shifts.

This discrepancy might also be associated with the relatively higher variety of the pulmonary patterns for the labels GGO and consolidation present in the various pulmonary diseases making up the non-Covid-19 group, as compared to the Covid-19 group. This group was, however, purposefully designed to resemble a broad range of non-Covid-associated pulmonary disease patterns, which represent Covid-19-mimics. Further, the large time frame in which these cases were collected, as well as a differing distribution amongst the three CT scanners used to generate these cases, might contribute to this finding.

Our experiments also show that our distance-based approach does not adequately detect poorly segmented cases for in-distribution data. This shortcoming reinforces the notion that uncertainty estimation methods, which are mainly designed to detect uncertain predictions in ID data, should complement OOD detection in practice. However, neither MC Dropout nor TTA were successful at assessing segmentation quality.

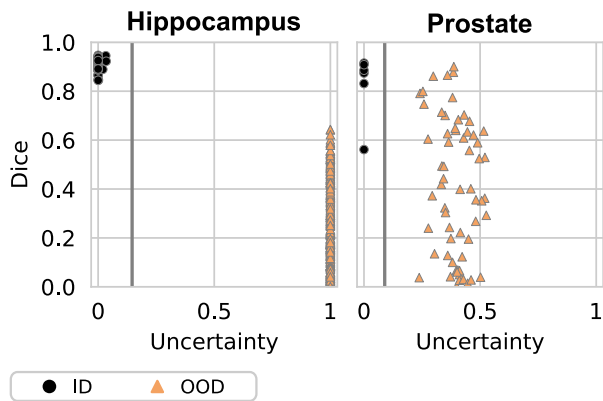


Fig. 14. Dice coefficient against normalised uncertainty for the segmentation of the hippocampus (left) and prostate (right) in MR images. Black circles are test ID (*MSD*) images, and orange triangles are OOD cases.

Our ablation study shows that intermediate network layers are the most informative for assessing distribution shifts. OOD samples do not display patterns that differ sufficiently from training samples in feature maps near the inputs or outputs of the model. In contrast, activations in intermediate layers allow the separation between ID and OOD cases. For the HighResNet model, which does not follow an encoder–decoder structure, dilated convolutions near the end of the model resulted in the best uncertainty estimates.

Finally, our *far-OOD* experiments on both CT and MR data confirm that our proposed method accurately detects cases very far from the training distribution. Such *far-OOD* cases may arise when an erroneous input is fed into the model, and automatically signalling such mistakes can be helpful for inexperienced users.

## 7. Conclusions

Despite ample progress in the development of segmentation solutions, these are not ready to be deployed in clinical practice. The main reason behind this is the fact that predictive models fail silently, coupled with a lack of appropriate quality controls to detect such behaviour. This is particularly true when it is not trivial to identify a faulty output, such as segmentation of SARS-CoV-2 lung lesions.

Increasingly, institutions are taking part in initiatives to gather large amounts of annotated, heterogeneous data and release it to the public. This could allow the training of robust models and potentially alleviate the burden of radiologists. However, even models trained with heterogeneous cohorts are susceptible to distribution shifts.

We propose a distance-based method to detect images far from the training distribution in a low-dimensional feature space, and find that this is a lightweight and flexible way to signal when a model prediction should not be trusted.

Future work should explore how to improve uncertainty calibration by identifying high-quality predictions. For now, our work increases clinicians' trust while translating trained neural networks from challenge participation to real clinics.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Most data used in this work is publicly available. We do not have permission to share the 100 in-house cases.

## Acknowledgements

This work was supported by the RACOON network under BMBF, Germany grant number [01KX2021]; and the Bundesministerium für Gesundheit (BMG), Germany with grant [ZMV11-2520DAT03A].

## References

- An, P., Xu, S., Harmon, S., Turkbey, E., Sanford, T., Amalou, A., Kassim, M., Varble, N., Blain, M., Anderson, V., et al., 2020. CT images in COVID-19. *Cancer Imaging Arch.*
- Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al., 2022. The medical segmentation decathlon. *Nat. Commun.* 13 (1), 1–13.
- Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D., 2019. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In: *International Conference on Learning Representations*.
- Bevandić, P., Krešo, I., Oršić, M., Šegvić, S., 2019. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In: *German Conference on Pattern Recognition*. Springer, pp. 33–47.
- Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K., 2015. NCI-ISBI 2013 challenge: automated segmentation of prostate structures. <http://dx.doi.org/10.7937/K9/TCIA.2015.zF0vIOPv>.
- Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network. In: *International Conference on Machine Learning*. PMLR, pp. 1613–1622.
- Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani, M., et al., 2015. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimer's Dement.* 11 (2), 175–183.
- Çalli, E., Murphy, K., Sogancioglu, E., van Ginneken, B., 2019. FRODO: Free rejection of out-of-distribution samples: application to chest x-ray analysis. In: *International Conference on Medical Imaging with Deep Learning—Extended Abstract Track*.
- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26 (6), 1045–1057.
- Fort, S., Ren, J., Lakshminarayanan, B., 2021. Exploring the limits of out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* 34.
- Fuchs, M., Gonzalez, C., Mukhopadhyay, A., 2021. Practical uncertainty quantification for brain tumor segmentation. In: *Medical Imaging with Deep Learning*.
- Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International Conference on Machine Learning*. PMLR, pp. 1050–1059.
- Golan, I., El-Yaniv, R., 2018. Deep anomaly detection using geometric transformations. *Adv. Neural Inf. Process. Syst.* 31.
- Gonzalez, C., Gotkowski, K., Bucher, A., Fischbach, R., Kaltenborn, I., Mukhopadhyay, A., 2021. Detecting when pre-trained nnu-net models fail silently for Covid-19 lung lesion segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 304–314.
- Gonzalez, C., Mukhopadhyay, A., 2021. Self-supervised out-of-distribution detection for cardiac CMR segmentation. In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*. In: *Proceedings of Machine Learning Research*, 143, PMLR, pp. 205–218, URL: <https://proceedings.mlr.press/v143/gonzalez21a.html>.
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: *International Conference on Machine Learning*. PMLR, pp. 1321–1330.
- Hein, M., Andriushchenko, M., Bitterwolf, J., 2019. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 41–50.
- Henderson, E., 2021. Leading pediatric hospital reveals top AI models in COVID-19 grand challenge. Accessed: 2021-02-28. <http://news-medical.net>.
- Hendrycks, D., Gimpel, K., 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: *International Conference on Learning Representations*.
- Hendrycks, D., Mazeika, M., Dietterich, T., 2018. Deep anomaly detection with outlier exposure. In: *International Conference on Learning Representations*.
- Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019. Using self-supervised learning can improve model robustness and uncertainty. *Adv. Neural Inf. Process. Syst.* 32.
- Hu, Y., Jacob, J., Parker, G.J., Hawkes, D.J., Hurst, J.R., Stoyanov, D., 2020. The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. *Nat. Mach. Intell.* 2 (6), 298–300.
- Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*. PMLR, pp. 448–456.
- Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18 (2), 203–211.

- Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Mingqing, Z., Xin, L., Xueyuan, D., Shucheng, C., Hao, W., Sen, M., Xiaoyu, Y., Ziwei, N., Chen, L., Lu, T., Yuntao, Z., Qiongie, Z., Guoqiang, D., Jian, H., 2020. COVID-19 CT lung and infection segmentation dataset. <http://dx.doi.org/10.5281/zenodo.3757476>.
- Jungo, A., Balsiger, F., Reyes, M., 2020. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Front. Neurosci.* 14, 282.
- Jungo, A., Reyes, M., 2019. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 48–56.
- Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* 30.
- Kohl, S.A., Romera-Paredes, B., Meyer, C., Fauw, J.D., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.A., Rezende, D.J., Ronneberger, O., 2018.
- Kondratyeva, E., Pominova, M., Popova, E., Sharaev, M., Bernstein, A., Burnaev, E., 2021. Domain shift in computer vision models for mri data analysis: an overview. In: *Thirteenth International Conference on Machine Vision*, Vol. 11605. SPIE, pp. 126–133.
- Kulaga-Yoskovitz, J., Bernhardt, B.C., Hong, S.-J., Mansi, T., Liang, K.E., Van Der Kouwe, A.J., Smallwood, J., Bernasconi, A., Bernasconi, N., 2015. Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset. *Sci. Data* 2 (1), 1–9.
- Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* 30, 6402–6413.
- Lee, K., Lee, H., Lee, K., Shin, J., 2018a. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: *International Conference on Learning Representations*.
- Lee, K., Lee, K., Lee, H., Shin, J., 2018b. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: *Advances in Neural Information Processing Systems*. pp. 7167–7177.
- Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F., 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput. Biol. Med.* 60, 8–31.
- Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In: *International Conference on Information Processing in Medical Imaging*. Springer, pp. 348–360.
- Liang, S., Li, Y., Srikant, R., 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In: *International Conference on Learning Representations*.
- Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Med. Image Anal.* 18 (2), 359–373.
- Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020a. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE Trans. Med. Imaging* 39 (9), 2713–2724.
- Liu, W., Wang, X., Owens, J., Li, Y., 2020b. Energy-based out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* 33, 21464–21475.
- Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* 39 (12), 3868–3878.
- Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B., 2020a. Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty. In: *Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., pp. 12756–12767.
- Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B., 2020b. Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty. *Adv. Neural Inf. Process. Syst.* 33, 12756–12767.
- Morozov, S., Andreychenko, A., Pavlov, N., Vladzimirskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I.A., Gelezhe, P., Gonchar, A., Chernina, V.Y., 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.
- Parekh, M., Donuru, A., Balasubramanya, R., Kapur, S., 2020. Review of the chest CT differential diagnosis of ground-glass opacities in the COVID era. *Radiology* 297 (3), E289–E302.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Louppe, G., 2012. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12.
- Pérez-García, F., Sparks, R., Ourselin, S., 2021. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput. Methods Programs Biomed.* 106236. <http://dx.doi.org/10.1016/j.cmpb.2021.106236>, URL: <https://www.sciencedirect.com/science/article/pii/S0169260721003102>.
- Pidhorskyi, S., Almohsen, R., Doretto, G., 2018. Generative probabilistic novelty detection with adversarial autoencoders. *Adv. Neural Inf. Process. Syst.* 31.
- Roefo, 2022. RACOON: das radiologische kooperative netzwerk zur beantwortung der großen fragen in der radiologie. <http://dx.doi.org/10.1055/a-1544-2240>, Accessed: 2022-03-08, <http://news-medical.net>.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.
- Roth, H., Xu, Z., Diez, C.T., Jacob, R.S., Zember, J., Molto, J., Li, W., Xu, S., Turkbey, B., Turkbey, E., et al., 2021. Rapid artificial intelligence solutions in a pandemic-the COVID-19-20 lung CT lesion segmentation challenge.
- Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.
- Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z., Mahapatra, D., 2021. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 226–238.
- Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE Trans. Med. Imaging* 36 (8), 1597–1606.
- Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338, 34–45.
- Wei, D., Zhou, B., Torrabi, A., Freeman, W., 2015. Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379*.
- Zakazov, I., Shirokikh, B., Chernyavskiy, A., Belyaev, M., 2021. Anatomy of domain shift impact on U-net layers in MRI segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 211–220.

---

## Contribution and impact

We significantly expand the validation of our method for this publication. For starters, besides the original Covid-19 experimental setup with three different datasets (which we now denominate *dataset shift*), we (a) explore a *diagnostic shift* scenario with non-Covid pneumonia subjects and (b) generate artificially transformed datasets across two types of shifts, namely applying affine operations and adding common CT artifacts, for three severities each. On this last data, we have total knowledge of the factors behind the shifts, which allows us to quantify how sensitive our method is to different changes in the data distribution.

Additionally, we look at two different MRI tasks, namely hippocampus and prostate segmentation; and validate our approach on a model that does not follow an encoder-decoder architecture. We also explore *far-OOD* cases to observe how our method reacts to inputs extremely far from the expected distribution to see whether it would accurately detect the erroneous use of the model. Specifically, we pass colon and spleen CTs through the model trained for lung lesion segmentation in chest CTs.

Our results show that the proposed method is effective across a number of problems and scenarios. The thorough evaluation provides insight into how different types of shifts affect model performance and how difficult these are to detect. We also confirm that the further data diverges from the training distribution, the easier these cases are to detect. This is observed in the gradually transformed dataset experiments, where the estimated uncertainty grows with the intensity of the transformation; and the far-OOD setting, which proved to be a very easy problem.

## Discussion and limitations

Beyond encouraging empirical results in identifying problematic OOD samples, the proposed method has several practical advantages. For starters, it works purely post-training, and the user only needs to modify the code base so as to extract network features during a forward pass. This makes the method flexible and applicable to complex image segmentation pipelines. It is also extremely lightweight, demanding no GPU resources other than to perform the forward pass that also segments the image. I, therefore, find it meaningful to perform this quality check in numerous settings.

One major limitation of the proposed approach is the simplification of assuming a multi-variate Gaussian distribution, which may not be the most suitable for describing the latent space. In particular, it fails to represent multi-modal data where there are several *clusters*. Training samples with certain characteristics may have successfully taught the model how to deal with similar cases, but may not be sufficiently represented so as to result in a low distance from the distribution mean.

Capturing complexities in the training features could also address one of the central practical concerns with the method, namely the fact that it is overly cautious in assessing the uncertainty for OOD predictions. Why is it that, in some cases, the model seems successful at extracting segmentation masks for features far from the training distribution? The obvious assumption would be that the distribution is not properly captured in our estimation.

In the future, we would like to develop strategies to identify *in what ways* a test sample is different from the training distribution. With this information, we could adapt the image or model in such a way that a high-quality prediction can be extracted.



---

## 3.2. Conclusions and outlook

---

Quantifying how suitable a model is for a particular image is key to ensuring the safe use of deep learning models in unpredictable clinical environments. We know that DNNs only produce meaningful outputs for inputs – and, in deeper layers, features – close to the training distribution, so samples should be flagged that do not meet this criterion.

I do not deem it necessary to adapt the architecture or training mechanism for this purpose. At the end of the day, the model will be used that obtains the best performance on the target task. Fortunately, methods exist that only look at the inputs or network features and do not require changing the architecture or even re-training the model. Calculating the distance to the training distribution in a down-sampled latent space is one strategy that reliably identifies both *far-OOD* cases where the model is used incorrectly and *near-OOD* images that contain artifacts or yet-unseen characteristics.

The difficulty with this approach lies in learning an expressive distribution that captures meaningful aspects of the data. It is also possible that learning *multiple* distributions for different sections of the latent space is more reasonable. For instance, a model trained with data from different sites may learn a specific set of features for each site. Then, it would be reasonable to calculate the distance to each such cluster.

Certain DNNs may additionally contain components that directly help assess their suitability. For instance, when deploying self-supervised models, the proxy loss can be monitored. A deterioration in the proxy task performance is often an indication of a similarly bad target task prediction.

Until now, I limited my work in this area to detecting *whether* a case is OOD. A more valuable objective may be to identify the *cause* behind this shift. Or, put in another manner, in what ways a test image is different from the training data. This information may help us select the best course of action for the low-quality prediction.

Imagine, for instance, a model that classifies between healthy subjects and patients suffering from different types of pneumonia from chest CTs. If the model were to receive a patient with a Covid-19 infection, which manifests differently from other conditions, the OOD detection system should signal that the case presents certain lesions it has not seen before. After several such cases, radiologists may start to note new disease patterns emerging. If, instead, the classifier were to receive the scan from a viral pneumonia patient acquired with a new scanner, it should inform the user that the image is different from the training base due to characteristics related to the acquisition. Here, domain adaptation approaches could potentially be used to obtain a successful prediction.

I hope to see more research in the coming years, preferably in the form of prospective studies in real clinics, on how OOD detection directs the use of further ML components.

---

## 4. Epistemic Uncertainty Estimation

---

Closely related to OOD detection is *uncertainty estimation*. Though the boundaries between both fields are blurry, uncertainty estimation performs best for in-distribution cases (Ovadia et al., 2019). We can distinguish between two types of uncertainty. *Aleatoric uncertainty* quantifies the inherent randomness in the prediction process. Imagine, for instance, the clearly irreducible uncertainty when a coin is tossed. The second is *epistemic uncertainty* which results from a lack of information or, from a more practical perspective, training data (Senge et al., 2014). Assuming that a model is suitable for the intended use, which we hope was proven by sufficient pre-market evaluation, we are interested in quantifying the epistemic uncertainty that arises from differences between a test image and the training data.

Uncertainty estimation methods typically obtain  $n$  different outputs  $\{y_i\}_{i \leq n}$  for the same input  $x$  under different conditions. These are acquired by training separate models for the same problem, or by using only one model but disturbing the input or the network features so that they produce different outputs (while still remaining within the expected range). We then calculate the variance between predictions to obtain an estimate of the uncertainty (Equation 4.1).

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2, \quad \mu = \frac{1}{n} \sum_{i=1}^n y_i \quad (4.1)$$

The intuition is that if the model – or model ensemble – is confident in the prediction for a particular input, small variations will not significantly affect it and the variance will be low. If, instead, predictions made by similar models or inputs are widely different, this could be an indication of high uncertainty. By averaging the predictions, we also typically obtain a better result (Lakshminarayanan et al., 2017), which is why ensembles are popular beyond their use for uncertainty estimation.

Two popular methods for estimating model uncertainty are *Monte Carlo (MC) Dropout* (Gal and Ghahramani, 2016) and *Deep Ensembles* (Lakshminarayanan et al., 2017). MC Dropout involves performing multiple forward passes through the model with activated Dropout layers (which are typically dormant at test time) and can be applied to any model that utilizes Dropout. Deep Ensembles, on the other hand, involves training multiple networks under slightly different conditions, such as from different initializations. These methods have been found to be effective in medical image segmentation (Jungo and Reyes, 2019; Jungo et al., 2020; Mehrtash et al., 2020), particularly Deep Ensembles. Another option is to slightly perturb inputs through *Test-Time Augmentation* (Ashukha et al., 2019), which can improve predictions for both single models and ensembles. Other methods, such as *Probabilistic Backpropagation* (Kohl et al., 2018), and *Stochastic Activation Pruning* (Monteiro et al., 2020), have shown better performance in some cases, but require specific training considerations, limiting their applicability.

Finally, *Bayesian Neural Networks (BNNs)* learn a *distribution* for each model parameter  $p(\theta|x, y)$  as opposed to a point estimate. A prior probability distribution is first defined over the parameters, which is updated during the training process to reflect the posterior. As  $p(y|x)$  is not tractable, *Variational*

---

*Inference (VI)* approximates the posterior by minimizing the Kullback-Leibler divergence between the approximating distribution and the true posterior. Typically, normal distributions are used, so a mean  $\mu$  and standard deviation  $\sigma$  are learned for each parameter  $\theta$ . Though VI takes an interesting perspective and allows us to quantify uncertainty for each parameter, it is computationally expensive and does not supply reliable uncertainties for complex tasks such as medical imaging problems.

In our recent work by Fuchs et al. (2022), we show that by combining VI with a multi-head ensemble mechanism and limiting the Bayesian layers to the deeper part of the network, we can leverage the advantages of both strategies while limiting the overhead. While the VI component captures the uncertainty within each local minimum, training multiple heads allows us to assess how different solutions to the problem diverge.

---

## 4.1. The paper: Improving robustness and calibration in ensembles with diversity regularization

---

Deep ensembles are highly popular as they produce reliable uncertainty scores and improve prediction quality with minimal additional work: models simply need to be replicated and perhaps trained from different initializations or training data subsets. Nevertheless, they also have several downsides. One is the additional computational overhead of training and performing inference with different models. There is also no guarantee that ensemble members will learn different functions or reach different local minima in the solution space, so each additional model may provide minimal additional value. Finally, the estimated uncertainties may only be meaningful for in-distribution data.

We attempt to solve these challenges with the paper *Improving robustness and calibration in ensembles with diversity regularization*, authored by Hendrik Mehtens, Anirban Mukhopadhyay and myself and presented at the 2022 DAGM German Conference on Pattern Recognition in Konstanz in September 30<sup>th</sup>, 2022. The publication summarizes the key findings of Mr. Mehtens' Master's thesis (Mehtens, 2021), which I had the pleasure to supervise. Mr. Mehtens presented the work and subsequently won the German Association for Pattern Recognition Young Researchers' Forum 2022 Best Master's Thesis award. We additionally uploaded a pre-print of the publication to arXiv on January 26<sup>th</sup>, 2022.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

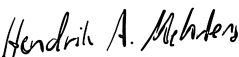


**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Improving robustness and calibration in ensembles with diversity regularization**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Improving robustness and calibration in ensembles with diversity regularization" (Mehrtens et al. 2022) was published as a full research paper at the "German Conference for Pattern Recognition (GCPR)". It constitutes a joint work of Hendrik Mehrtens, Camila González and Anirban Mukhopadhyay.*

*As corresponding and leading author, H. Mehrtens led the overall research design, literature review and writing process of the paper. The choice of methodological framework and experimental setup were done by C. González and H. Mehrtens together. H. Mehrtens implemented the code and conducted the experiments. H. Mehrtens and C. González contributed to the analysis of the data and results. The methodology, results and discussion were mainly written by H. Mehrtens, and C. González continuously reviewed the content and supplied feedback. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor of this work, who also contributed with continuous feedback during all phases of the paper writing process. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum:	<u>01 / 10 / 2023</u>	<u>01 / 09 / 2023</u>	<u>01 / 16 / 2023</u>
Unterschrift:	<u></u>	<u></u>	<u></u>
	Hendrik Mehrtens	Camila González	Anirban Mukhopadhyay

# Improving Robustness and Calibration in Ensembles with Diversity Regularization

Hendrik Alexander Mehrtens<sup>1,2</sup>[0000-0003-1234-5041], Camila Gonzalez<sup>1</sup>[0000-0002-4510-7309], and Anirban Mukhopadhyay<sup>1</sup>[0000-0003-0669-4018]

<sup>1</sup> Technische Universität Darmstadt, 64289 Darmstadt, Germany  
{anirban.mukhopadhyay, camila.gonzalez}@gris.tu-darmstadt.de  
<sup>2</sup> Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany  
hendrikalexander.mehrtens@dkfz.de

**Abstract.** Calibration and uncertainty estimation are crucial topics in high-risk environments. Following the recent interest in the diversity of ensembles, we systematically evaluate the viability of explicitly regularizing ensemble diversity to improve robustness and calibration on in-distribution data as well as under dataset shift. We introduce a new diversity regularizer for classification tasks that uses out-of-distribution samples and increases the overall accuracy, calibration and out-of-distribution detection capabilities of ensembles. We demonstrate that diversity regularization is highly beneficial in architectures where weights are partially shared between the individual members and even allows to use fewer ensemble members to reach the same level of robustness. Experiments on CIFAR-10, CIFAR-100, and SVHN show that regularizing diversity can have a significant impact on calibration and robustness, as well as out-of-distribution detection.

**Keywords:** diversity · ensembles · robustness · calibration

## 1 Introduction

When a machine learning system is used in high-risk environments, such as medicine and autonomous driving, a well-calibrated estimate of the uncertainty is necessary. A model is said to be *calibrated* [9] if the confidence of its predictions reflects its true probability of being correct. However, deep neural networks tend to be overconfident in their predictions [9] leading to multiple recent approaches attempting to improve their calibration [2,32]. Furthermore, models need to be robust to shifts in the data domain, which can for example arise in the data shift between the training and deployment domains.

To this day, Deep Ensembles [21] outperform most other approaches. A common explanation for the improved performance is the high diversity of solutions in the ensemble [4,6,27], which is mostly generated by training from different parameter initializations. While this approach works well empirically, distance in parameter space generated through training from different starting positions

does not guarantee diversity in the solution space, which we refer to as *functional diversity* [43]. However, ensuring a diverse set of solutions in an ensemble is critical to its performance [6,43].

Following recent interest in the topic of diversity in neural network ensembles [6], many publications try to implicitly generate diversity by training with different architectures [1,45], different data augmentations [38] and different hyperparameters [42]. However, this approach to generate diversity is sub-optimal, as it does not guarantee diversity. Additionally, choosing the right architectures and hyperparameters requires a lot of design decisions and is thereby time-consuming.

On the other side, functional diversity can be regularized explicitly [27], an idea recently used to improve adversarial robustness in ensembles [16,33]. Although these explicit approaches guarantee diversity of predictions, they rely on diversity measures on the original training data, which can lead to a degradation in accuracy. Additionally, these approaches do not perform well in tasks of out-of-distribution detection and the naive implementation requires the simultaneous training of multiple ensemble members, which is expensive and can be prohibitive in some tasks.

In our experiments, we put a special focus on ensembles that share parameters between the members. While these architectures require much less computational time, the lower ratio of independent parameters per member leads to a reduction of diverse predictions [25], which naturally lends itself to using explicit diversity maximization. For this, we use ensemble architectures with an increasing ratio of shared parameters between members and show that the effect of diversity regularization on robustness and calibration increases with a higher ratio of shared parameters.

We introduce the **Sample Diversity regularizer (SD)** that instead of using in-distribution images to diversify the predictions, uses out-of-distribution images. This, as we show, can be sampled from noise, not requiring an external dataset and increases accuracy and calibration under dataset shift, while also increasing the out-of-distribution detection capabilities of the model, contrary to our other baseline regularizers. The proposed regularizer can also be combined for greater effect with the other explicit diversity regularizers. Taking inspiration from the methods of Shui et al. [35], we systematically evaluate the effectiveness of explicit diversity regularization, coming to the conclusion that diversity regularization is especially useful when encountering dataset shift [32], even reducing the number of ensemble members needed for the same performance and allowing for the training of light-weight approximate ensemble architectures instead of full ensembles.

To summarize, our contributions are as follows:

- We demonstrate that diversity regularization is **highly effective for architectures with a high ratio of shared parameters**, reducing the number of needed ensemble members under dataset shift and allowing for smaller architectures.

- We introduce the **Sample Diversity regularizer**, which **increases the accuracy and calibration under dataset shift, as well as the out-of-distribution detection capabilities** and can be combined with existing diversity regularizers for greater effect.

## 2 Related work

In recent years, calibration of deep neural networks has become a focus in machine learning research. Although multiple approaches, from temperature scaling [9], MC Dropout [7,18] to Variational Inference methods [3,28] have been explored, neural network ensembles have demonstrated that they produce the best-calibrated uncertainty estimates [2,32,21].

An important property of well-calibrated models is whether they still give reasonable uncertainties when encountering dataset shift, as this setting better reflects real-world conditions. Ovadia et al. [32] compared multiple approaches using the CIFAR-10-C, CIFAR-100-C and ImageNet-C datasets by Hendrycks et al. [12], coming to the conclusion that Deep Ensembles [22] outperformed every other approach, making them the de-facto standard for uncertainty estimation and robustness.

The superiority of ensembles in these task has been partly attributed to the diversity between the individual members [6]. Ensemble diversity, in general, has long been a research topic in machine learning with many early works recognizing it as a key principle in the performance of ensembles [4,27]. Recently, a greater focus has been placed on improving diversity in neural network ensembles by different *implicit* means, for example by providing each ensemble member with differently augmented inputs [38], building ensembles out of different neural network architectures [1,45,36] or training ensemble members with different hyperparameters [42].

*Explicit* approaches on the other hand try to maximize the diversity between ensemble members by orthogonalizing their gradients [16], decorrelating their predictions on all classes [27,35] or on randomly sampled noise inputs [15] or orthogonalizing only on non-correct classes [33]. Another strategy is to increase diversity in the internal activations or parameters of the ensemble members [34,37,23], which forms a promising direction but requires computationally expensive adversarial setups. Finally, there are sampling-based methods that try to maximize the diversity of the sampling procedure, for example through Determinantal Point Processes [20,40]. The advantage of these explicit approaches is that they can directly control the diversity in the ensemble and do not rely on decisions with indirect and often unclear consequences.

As training ensembles is expensive, multiple methods have tried to reduce training costs. Snapshot-based methods [8,14] save multiple epochs along a training trajectory, Batch Ensembles [41] generate individual ensemble members by addition of a per-member Rank-1 Hadamard-product and TreeNets [25] approximate a Deep Ensemble by sharing the lower levels of a network between members. Furthermore, distillation approaches were proposed [39,44] that try to compress

multiple networks into a single one. However, these approaches tend to reduce the diversity between the individual members, by either sharing parameters between them or not training them independently, leading to a reduction in accuracy and calibration.

In this work we show that diversity regularization is highly useful in parameter shared ensembles and that diversity regularization can not only help with accuracy and under dataset shift but also with out-of-distribution detection. Taking inspiration from Jain et al. [15] we introduce an explicit diversity regularizer for classification that uses out-of-distribution samples, leaving the predictions on the original data intact.

### 3 Methods and metrics

For our evaluation, we consider a classification task with  $C$  classes. Given a data point  $x \in \mathbb{R}^L$  out of a dataset with  $N$  entries and its corresponding one-hot label  $\hat{y} \in \mathbb{R}^C$ , the prediction of the  $j$ -th member of an ensemble with  $M$  members is called  $f(x, \theta_j) = y_j$ , where  $\theta_j \in \mathbb{R}^P$  are the parameters of the  $j$ th ensemble member. We refer to the mean of all predictions as  $\bar{y}$ .

In this section, we describe the evaluated regularization functions, architectures, and metrics as well as introduce our novel approach to diversity regularization.

#### 3.1 Regularizers

Given an image  $x$ , a label  $\hat{y}$  and the ensemble predictions  $y_i, i \in [1, \dots, M]$ , all regularizers  $\mathcal{L}_{reg}$ , which will be introduced in the following paragraphs, work as a regularizer to the cross-entropy ( $CE$ ) loss, where  $\lambda_{reg}$  is a hyper-parameter that is chosen for each individual method.

$$\mathcal{L}_{total}(\hat{y}, y_1, \dots, y_M) = \mathcal{L}_{CE}(\hat{y}, \bar{y}) - \lambda_{reg} \mathcal{L}_{reg}(\dots) \quad (1)$$

For our experiments, we select a set of regularization functions that compute a measure of similarity of the individual ensemble members' predictions. An illustration of the general structure can be seen in Figure 1.

Regularizers under consideration are our *Sample Diversity* regularizer, the *ADP* [33] regularizer, which was recently introduced for increasing robustness in ensembles to adversarial attacks, and the *Negative Correlation* regularizer.

Additionally we consider the average pair-wise  $\chi^2$  distance (see Eq. 4). All these regularizers encourage the individual members to have diverse predictions given the same input and can therefore be seen as increasing the functional diversity.

**Negative Correlation:** The *Negative Correlation* regularizer was first used by Liu et al. [27] to increase the diversity in neural network ensembles. The key insight was that the error of an ensemble depends upon the correlation of the errors between individual members [4]. Originally designed for regression tasks,



it was already used by Shui et al. [35] to improve the diversity and calibration in neural network ensembles in classification tasks. This approach however reduces the accuracy of the ensemble and can lead to training instabilities.

$$NegCorr(y_1, \dots, y_M) = - \sum_i^C ((y_i - \bar{y}) \cdot (\sum_{i \neq j} y_j - \bar{y})) \quad (2)$$

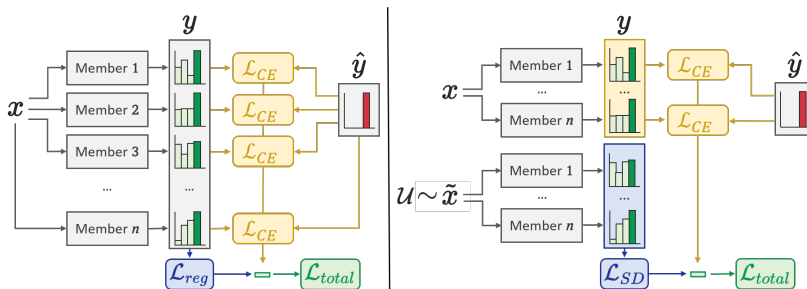
**ADP:** The *ADP* regularizer [33] orthogonalizes the predictions of the ensemble members on the non-correct classes during training.

Given a correct class  $k$ , the vector of the predictions for the non-correct classes are formed  $y_i^{\setminus k} = (y_i^1, \dots, y_i^{k-1}, y_i^{k+1}, \dots, y_i^C)$ , re-normalized and stacked into a matrix  $Y_{\setminus k} \in \mathbb{R}^{(C-1) \times M}$ . Furthermore an entropy regularizer ( $H$ ) is used to prevent extreme solutions. Together the regularizer is optimized using the hyperparameters  $\alpha$  and  $\beta$ .

$$ADP(\bar{y}^{\setminus k}, y_1^{\setminus k}, \dots, y_C^{\setminus k}) = \alpha \cdot H(\bar{y}^{\setminus k}) + \beta \log(\det(Y_{\setminus k}^T \cdot Y_{\setminus k})) \quad (3)$$

$\chi^2$  **distance:** As a distance measure between distributions, we implement the average pair-wise  $\chi^2$  distance between the members' predictive distributions as a regularizer. Like the likelihood, the measure lives on the range  $[0, 1]$  and the regularizer can be computed as

$$\chi^2(y_1, \dots, y_M) = \log \left( \frac{1}{M \cdot (M - 1)} \sum_{i \neq j} \sum_{k=1}^C \frac{y_i^{(k)} - y_j^{(k)}}{y_i^{(k)} + y_j^{(k)}} \right) \quad (4)$$



**Fig. 1.** Conceptual figure showcasing the differences between approaches. (left) Given an input  $x$ , the individual members are not only optimized individually with regard to the cross-entropy loss but the predictions are additionally regularized by a diversity regularizer ( $L_{reg}$ ). The predictions of the individual members on the original input are compared by a diversity regularizer. (right) Our *Sample Diversity* ( $L_{SD}$ ) approach utilizes additional inputs, sampled from the uniform distribution, to compute a measure of diversity as a regularizer. This preserves the original predictions on the training data.

### Sample Diversity:

Building on the work of Jain et al. [15] and the loss-formulation of *ADP* [33], we introduce a similar regularizer for classification tasks, which we illustrate in Figure 1. Instead of regularizing diversity on the predictions of in-distribution data points, which could degrade performance, we generate out-of-distribution data points and enforce predictive orthogonality there. The loss reaches a minimum if all predictions are orthogonal on the sampled data points and thereby diverse but correct on the in-distribution data. Image batches are sampled from the uniform probability distribution (white noise). Given all logits outputs for a sampled data point  $\tilde{x} \sim U^{H \times W}$ , of all  $M$  ensemble members  $(\tilde{y}_1, \dots, \tilde{y}_M)$ , normalized to length one and stacked in a matrix  $\tilde{Y} \in \mathbb{R}^{C \times M}$ , we maximize Eq. 5 as our regularizer.

$$\text{SampleDiversity}(\tilde{y}_1, \dots, \tilde{y}_M) = \log(\det(\tilde{Y}^T \cdot \tilde{Y})) \quad (5)$$

Other possible formulations are evaluated in the supplemental material, for comparability we stick to the *ADP* loss formulation. In the out-of-distribution detection literature, multiple other approaches that utilize OOD data during training exist, however these approaches act on single neural networks, utilize adversarial generators and experiment in the out-of-distribution detection domain [29,24,13]. Our goal is to formulate a practical functional diversity regularizer that utilizes the strength of ensembles, while not requiring expensive adversarial training.

### 3.2 Architectures

As more shared parameters reduce the computational resources required when training an ensemble but also the diversity of the ensemble, we study if higher dependency between members, increases the viability of diversity regularization. To this end, we compare the independently trained Deep Ensembles of randomly initialized neural networks [21] without adversarial training with TreeNets [25] that approximate a Deep Ensemble by sharing a base part of the network with each member, as well as Batch Ensembles [41] that generate their members by adding a Rank-1 Hadamard product to the parameter matrices of a base network and have the least number of independent parameters. We limit the scope of our study to the aforementioned architectures, although other architectures like the MiMo architecture [10] exist, as they are closest in structure to the Deep Ensemble.

### 3.3 Metrics

When working with calibration it is not only important to be well-calibrated on the original data but also under reasonable dataset shifts, which is crucial for real-world application. To evaluate this, corrupted datasets are used that simulate realistic noise and corruptions settings. All our metrics will be reported on the original datasets, as well as under dataset shift. In addition to accuracy

and negative log-likelihood ( $NLL$ ), we measure additional metrics, which are explained in the following:

**Calibration:** A commonly used measure of calibration is the *Expected Calibration Error (ECE)* [30]. As noticed by Ashuka et al. [2] this metric may not produce consistent rankings between models. For this reason, temperature scaling [9] with five-fold cross-validation on the test-set is deployed to generate consistent results. The temperature is computed for each dataset on the uncorrupted version. Our scores are computed after applying temperature scaling to the predictions. The temperature is chosen to minimize the negative log-likelihood, as proposed by Guo et al. [9].

**AUC-ROC:** The ability of detecting out-of-distribution data is tested, as intuitively more diverse ensemble members should produce more diverse predictions when evaluated on out-of-distribution data. We use the confidence of the average prediction of the ensemble as threshold classifier for distinguishing between in-distribution ( $ID$ ) and out-of-distribution ( $OOD$ ) data. Following [2] the AUC-ROC metric is reported for  $OOD$  detection.

**Table 1.** Experiments on CIFAR-10. Comparison of diversity regularization on different architectures with ensemble size 5 under dataset shift on the original (org.) data and highest corruption level (corr.).

Model	Method	Accuracy $\uparrow$		ECE $\downarrow$	
		org.	corr.	org.	corr.
DeepEns.	ind.	<b>.936</b> $\pm$ .001	.543 $\pm$ .010	.023 $\pm$ .001	.170 $\pm$ .014
	ADP	.933 $\pm$ .000	.549 $\pm$ .005	.032 $\pm$ .002	<b>.126</b> $\pm$ .010
	NegCorr.	.934 $\pm$ .001	.538 $\pm$ .002	.023 $\pm$ .001	.164 $\pm$ .007
	$\chi^2$	.934 $\pm$ .001	.542 $\pm$ .006	.023 $\pm$ .000	.171 $\pm$ .008
	<b>SampleDiv.</b>	.933 $\pm$ .001	<b>.579</b> $\pm$ .004	<b>.022</b> $\pm$ .001	.134 $\pm$ .007
TreeNet	ind.	.919 $\pm$ .002	.523 $\pm$ .01	.035 $\pm$ .001	.234 $\pm$ .010
	ADP	.917 $\pm$ .002	.535 $\pm$ .019	<b>.024</b> $\pm$ .000	<b>.180</b> $\pm$ .031
	NegCorr.	.918 $\pm$ .003	.528 $\pm$ .013	.027 $\pm$ .002	.200 $\pm$ .014
	$\chi^2$	<b>.920</b> $\pm$ .004	.517 $\pm$ .013	.027 $\pm$ .001	.238 $\pm$ .013
	<b>SampleDiv.</b>	.916 $\pm$ .002	<b>.545</b> $\pm$ .007	.030 $\pm$ .002	.213 $\pm$ .014
BatchEns.	ind.	.905 $\pm$ .001	.512 $\pm$ .019	.097 $\pm$ .002	.285 $\pm$ .014
	ADP	<b>.906</b> $\pm$ .002	.517 $\pm$ .011	<b>.032</b> $\pm$ .008	<b>.171</b> $\pm$ .049
	NegCorr.	.904 $\pm$ .001	.503 $\pm$ .002	.072 $\pm$ .021	.258 $\pm$ .030
	$\chi^2$	.905 $\pm$ .002	.503 $\pm$ .014	.058 $\pm$ .007	.265 $\pm$ .030
	<b>SampleDiv.</b>	.904 $\pm$ .000	<b>.545</b> $\pm$ .007	.037 $\pm$ .015	.175 $\pm$ .032

## 4 Experiments and results

We first describe the general setup that is used in all of our experiments. After that, we test the effect of our different diversity regularizers on the accuracy, *NLL* and calibration and later focus on out-of-distribution detection.

### 4.1 Datasets, models, and training

The base architecture for all our experiments is a ResNet-20 [11]. We train our models on the CIFAR-10, CIFAR-100 [19] and SVHN [31] datasets. For experiments under dataset shift, we use the corrupted versions of the CIFAR-10 and CIFAR-100 datasets created by Hendrycks et al. [12] and additionally create a corrupted version of the SVHN dataset using all 19 corruptions with 5 levels of corruption intensity.

All experiments are conducted, unless otherwise stated, with a learning rate of  $1e-4$ , a  $L_2$  weight decay of  $2e-4$ , a batch size of 128 and Adam [17] as the optimizer, with the default  $\beta_1$  and  $\beta_2$  parameters. Each model is trained for 320 epochs. For augmentation, we use random crops and random horizontal flips, as described by Kaiming et al. [11]. The optimal temperatures are computed by five-fold cross-validation on the test dataset, as suggested by Ashukha et al. [2].

When using the TreeNet architecture the ResNet is split after the second pooling operation. The cross-entropy loss is computed for each member individually and then combined. When training the Batch Ensemble, each member is trained with the same inputs at each step, so it is possible to compare the predictions of the individual members. Batch Ensemble was originally trained by splitting a batch over the ensemble members in each step. When evaluating the impact of this change, we found no significant differences between the two training methods. The comparison can be found in the supplemental material.

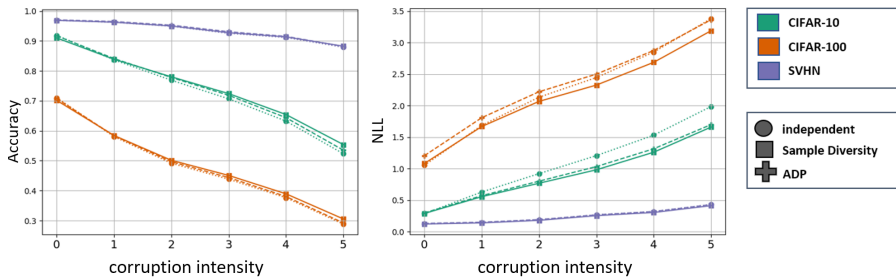
Each experiment is performed 3 times and we report the mean performance together with the standard deviation. Whenever possible, hyperparameters are chosen as presented in the original papers. All other parameters were fine-tuned by hand on a 10% split of the training data.

When training with the *ADP* regularizer, we use the parameters  $\alpha = 0.125$ ,  $\beta = 0.5$  which performed best for us in preliminary experiments. Those are the original parameters reported in the paper scaled by a factor of 0.25. For the *Sample Diversity* regularizer, we choose the number of sampled images equal to the original batch size. The images are sampled uniformly on all 3 channels in the range  $[0, 1]$ . We then choose  $\lambda_{SD} = 0.5$  for training. The  $\chi^2$  baseline used  $\lambda_{\chi^2} = 0.25$ . The *Negative Correlation* regularizer proved hard to train in a stable manner. We use  $\lambda_{NC} = 1e-5$ , as values above this threshold destabilized the training process.

### 4.2 Diversity regularization under dataset shift

We train the Deep Ensemble, TreeNet, and Batch Ensemble architectures on CIFAR-10, CIFAR-100, and SVHN. The experiments are performed with 5 ensemble members. On all three datasets, we compare the independently trained

ensembles, which we refer to as 'ind.' in our figures, with the regularized variants. We then evaluate all models on the corrupted versions of the datasets, comparing the accuracy,  $NLL$  and  $ECE$ .

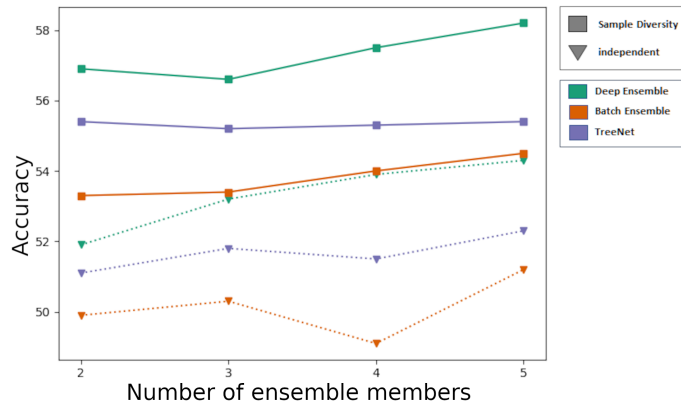


**Fig. 2.** Accuracy (left) and  $NLL$  (right) with different regularizers over different datasets.

Table 1 shows the results of our experiments on CIFAR-10 and the corrupted variant with all architectures. We compare the accuracy and  $ECE$  on the original data and on the highest corruption level. The results for the CIFAR-100 and SVHN datasets can be found in the supplemental material.

The *Sample Diversity* regularizer outperforms all other regularization functions in terms of accuracy on the corrupted data, improving the accuracy under dataset shift by 3.6% (Deep Ensemble), 2.3% (TreeNet) and 3.3% (Batch Ensemble), as can be seen on all architectures under dataset shift. Both the *Sample Diversity* and *ADP* regularizer outperform the other approaches in terms of  $ECE$ . The only exception occurs on the non-corrupted data with the Deep Ensembles architecture, where the *ADP* regularizer slightly decreases the calibration. Overall the  $\chi^2$  and *Negative Correlation* regularizer perform worse. This is most likely due to the fact that the diversity in these regularizers is also enforced on the correct class. When training these regularizers we also observed training instabilities.

As hypothesized the diversity regularization is effective when using constrained ensemble architectures. This is particularly noticeable for the Batch Ensemble architecture, which has the highest amount of shared weights per member, but also on the TreeNet architecture, a significant decrease of the  $ECE$  is observable, even on the original data, compared to the Deep Ensemble architecture, where diversity regularization performs worse. An interesting observation is that the TreeNet and Batch Ensemble regularized with the *Sample Diversity* loss outperform the Deep Ensemble of the same size (54.5% on both architectures, compared to 54.3% for the unregularized Deep Ensemble) on the corrupted data in terms of classification accuracy. Looking at the results it is clear that regularizing diversity helps in improving robustness to dataset shifts. It improves ensemble calibration, lowering the  $ECE$  under dataset shift significantly. The

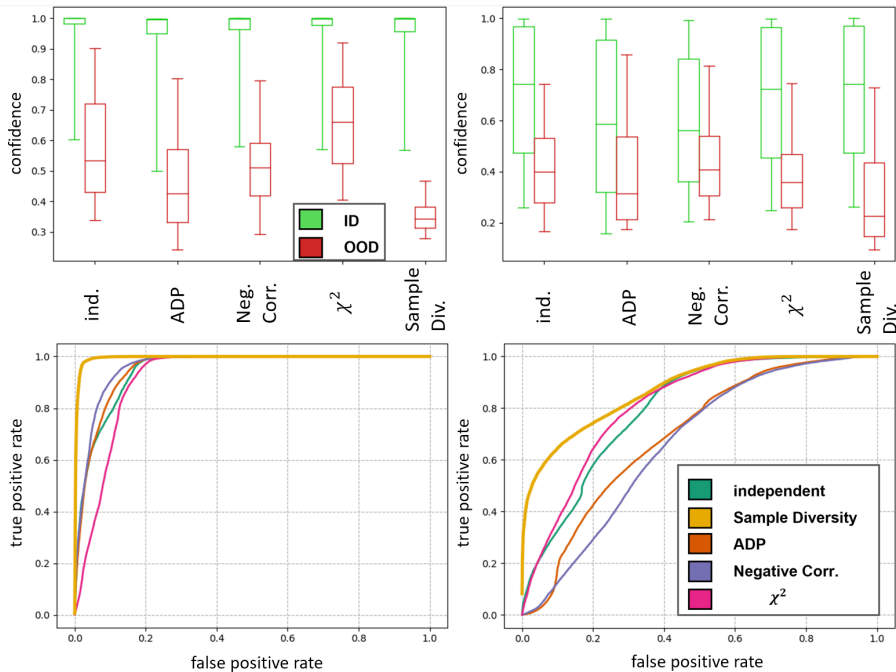


**Fig. 3.** Comparison of the three different architectures over the ensemble sizes 2 to 5 on the highest corruption level on CIFAR-10.

displayed metrics show a clear split between *ADP* and *Sample Diversity* on one side and the normal training routine on the other side.

Figure 2 compares the mean accuracy and negative log-likelihood of the *Sample Diversity* and *ADP* regularizer over all three datasets. We use a TreeNet with 5 members. The x-axis denotes the corruption level, the colors encode the dataset, while the line style and marker encode the regularizer. The *Sample Diversity* regularizer (solid, square) consistently improves the accuracy and decreases the negative log-likelihood under dataset shift. This difference is especially noticeable for the CIFAR-10 and CIFAR-100 datasets, on SVHN all methods stay relatively close to each other. The *ADP* (dashed, plus) regularizer on the other hand can even strongly decrease the negative log-likelihood on the original data, as can be seen with the CIFAR-100 results.

Figure 3 compares the effectiveness of *Sample Diversity* regularization on the highest corruption level of CIFAR-10 over the different ensemble sizes 2 to 5, comparing the mean accuracy, *NLL* and *ECE*. The colors encode the architecture, while the line style and marker encode the regularizer (*Sample Diversity* or independent training). As can be seen in the figure, diversity regularization is even highly effective when using as few as 2 ensemble members and does not require a large pool of members. Even a TreeNet or BatchEnsemble with 2 members, outperforms the unregularized equivalent with 5 members. This strongly reduces the number of ensemble members required for the same performance and shows that even lightweight ensemble architectures can outperform a Deep Ensemble. A table with detailed results can be found in the supplemental material.



**Fig. 4.** Distribution of confidence (top) and ROC curve for distinguishing between ImageNet and ID data (bottom) for a TreeNet architecture with 5 members with different regularization’s on CIFAR-10 (left) and CIFAR-100 (right).

### 4.3 Out-of-distribution detection

Figure 4 shows the distribution of confidence and the receiver operating characteristic (*ROC*) for differentiating between in-distribution and out-of-distribution data, which was in this experiment chosen as tinyImageNet [26], a 200-class subset of the ImageNet [5] dataset. The evaluated models are TreeNet architectures with 5 members, trained on CIFAR-10 and CIFAR-100. We do not report on SVHN, as every model there reached a near-perfect separation between in- and out-of-distribution data.

Looking at the ROC on CIFAR-10, *Negative Correlation* and *ADP* improve the separation slightly, while *Sample Diversity* strongly increases the dataset separation. Things look different on CIFAR-100 where all regularizers but *Sample Diversity* decrease the separation of the two datasets compared to independent training. To explain this, we take a look at the differences in the confidence distributions across the datasets. While all models are fairly confident in their predictions on CIFAR-10, most likely due to the few well-separated classes, on CIFAR-100 every model is highly uncertain, with confidence distributions between in-distribution data and OOD highly overlapping. There, the increased uncertainty that *Negative Correlation* and *ADP* introduce on in-distribution

**Table 2.** AUC-ROC over three runs, on separating in-distribution data and out-of-distribution data. Entries marked with '-' diverged.

Model (trained on)	Method	AUC-ROC $\uparrow$	
		CIFAR-10	CIFAR-100
DeepEns.	indi.	.980 $\pm$ .014	.798 $\pm$ .025
	ADP	.965 $\pm$ .017	.804 $\pm$ .034
	NCL	<b>.993</b> $\pm$ .001	.729 $\pm$ .038
	$\chi^2$	.983 $\pm$ .011	.834 $\pm$ .037
	<b>SD</b>	.982 $\pm$ .012	<b>.919</b> $\pm$ .026
TreeNet	indi.	.947 $\pm$ .044	.799 $\pm$ .049
	ADP	.952 $\pm$ .008	.695 $\pm$ .050
	NCL	.960 $\pm$ .011	.663 $\pm$ .097
	$\chi^2$	.916 $\pm$ .020	.815 $\pm$ .019
	<b>SD</b>	<b>.995</b> $\pm$ .003	<b>.877</b> $\pm$ .122
BatchEns.	indi.	.928 $\pm$ .008	.497 $\pm$ .187
	ADP	.909 $\pm$ .026	.595 $\pm$ .110
	NCL	.934 $\pm$ .076	-
	$\chi^2$	.974 $\pm$ .008	<b>.809</b> $\pm$ .122
	<b>SD</b>	<b>.991</b> $\pm$ .004	.614 $\pm$ .080

predictions is a disadvantage as the confidence distributions now tend to overlap more. On the other hand, *Sample Diversity* that only encourages orthogonality on OOD data improves the OOD detection capability.

Table 2 reports the AUC-ROC, as suggested by Ashuka et al. [2], for all our evaluated models. None of the baseline regularizers is able to consistently increase the AUC-ROC over the level reached by independent training. *Sample Diversity*, even though it is not the best in every single experiment, outperforms independent training and nearly all other regularizers consistently in every setting. We conclude that *Sample Diversity* can not only increase the robustness and calibration but at the same time also the out-of-distribution detection capabilities of the model, while only requiring uninformative out-of-distribution noise and no additional datasets.

Ablation studies, as well as more detailed results can be found in the supplemental material. There the experiments indicate that the *ADP* loss formulation is sub-optimal and future research could increase the viability of diversity regularization further. Furthermore, adversarial training for the image generation of *Sample Diversity* or using real out-of-distribution data like ImageNet instead of the uniformly sampled noise images, does not lead to any further improvements. We also conduct experiments, measuring the impact of distance in parameter space and the different behaviors of our tested regularizers. Finally, we test other architectures are tested to confirm our results.



## 5 Conclusion

We conduct a comprehensive study comparing different popular diversity regularization methods on robustness, calibration and out-of-distribution detection benchmarks, over multiple datasets and architectures. Furthermore, we introduce the *Sample Diversity* regularizer, which is well suited for **improving accuracy and ECE and can be combined with the ADP regularizer for greater effect**. Contrary to other regularizers, our regularizer **also increases the out-of-distribution detection capabilities**. Our experiments show that **diversity regularized ensembles are better in terms of accuracy and calibration under dataset shift**. Regularizing ensembles beyond the diversity reached by independent training especially on architectures with shared parameters is beneficial. Even **the TreeNet and Batch Ensemble can outperform a Deep Ensemble in terms of robustness to dataset shift when diversity regularization is used**, even when we use fewer members.

## Acknowledgements

I would like to thank the Deutsches Krebsforschungszentrum (DKFZ) for supporting the publication of this paper. I am grateful for everyone who proof-read the present and earlier versions of the manuscript, providing helpful insights.

## References

1. Antorán, J., Allingham, J.U., Hernández-Lobato, J.M.: Depth uncertainty in neural networks. In: Advances in neural information processing systems (2020)
2. Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.: Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In: International Conference on Learning Representations (2020)
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural networks. In: In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015) (2015)
4. Brown, G.: Diversity in Neural Network Ensembles. Ph.D. thesis, University of Manchester (03 2004)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Fort, S., Hu, H., Lakshminarayanan, B.: Deep ensembles: A loss landscape perspective (2019)
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. pp. 1050–1059 (2016)
8. Garipov, T., Izmailov, P., Podoprikin, D., Vetrov, D., Wilson, A.G.: Loss surfaces, mode connectivity, and fast ensembling of dnns. In: Advances in neural information processing systems (2018)

9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (06–11 Aug 2017)
10. Havasi, M., Jenatton, R., Fort, S., Liu, J.Z., Snoek, J., Lakshminarayanan, B., Dai, A.M., Health, G., Tran, D.: Training independent subnetworks for robust prediction
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
12. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
13. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations (2019)
14. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get m for free (2017)
15. Jain, S., Liu, G., Mueller, J., Gifford, D.: Maximizing overall diversity for improved uncertainty estimates in deep ensembles. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4264–4271 (2020)
16. Kariyappa, S., Qureshi, M.K.: Improving adversarial robustness of ensembles with diversity training (2019)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
18. Kingma, D.P., Salimans, T., Welling, M.: Variational dropout and the local reparameterization trick. In: Advances in neural information processing systems. pp. 2575–2583 (2015)
19. Krizhevsky, A.: Learning multiple layers of features from tiny images. (2009)
20. Kulesza, A.: Determinantal point processes for machine learning. In: Foundations and Trends® in Machine Learning. vol. 5, p. 123–286. Now Publishers (2012)
21. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (2017)
22. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
23. Larrazabal, A.J., Martínez, C., Dolz, J., Ferrante, E.: Orthogonal ensemble networks for biomedical image segmentation. In: MICCAI (2021)
24. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: International Conference on Learning Representations (2018)
25. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D.: Why m heads are better than one: Training a diverse ensemble of deep networks (2015)
26. Li, F.F., Karpathy, A., Johnson, J.: The tinyimagenet dataset - kaggle, <https://www.kaggle.com/c/tiny-imagenet>
27. Liu, Y., Yao, X.: Ensemble learning via negative correlation. In: Neural Networks. vol. 12 (1999)

28. Maddox, W.J., Izmailov, P., Garipov, T., Vetrov, D.P., Wilson, A.G.: A simple baseline for bayesian uncertainty in deep learning. In: *Advances in Neural Information Processing Systems*. pp. 13153–13164 (2019)
29. Malinin, A., Gales, M.: Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In: *Advances in neural information processing systems* (2019)
30. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: *AAAI*. p. 2901–2907. *AAAI'15*, AAAI Press (2015)
31. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011* (2009)
32. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: *Advances in Neural Information Processing Systems*. pp. 13991–14002 (2019)
33. Pang, T., Xu, K., Du, C., Chen, N., Zhu, J.: Improving adversarial robustness via promoting ensemble diversity. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 97, pp. 4970–4979. PMLR (09–15 Jun 2019)
34. Rame, A., Cord, M.: Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. In: *International Conference on Learning Representations* (2021)
35. Shui, C., Mozafari, A.S., Marek, J., Hedhli, I., Gagné, C.: Diversity regularization in deep ensembles. In: *International conference on learning representations* (2018)
36. Singh, S., Hoiem, D., Forsyth, D.: Swapout: Learning an ensemble of deep architectures (2016)
37. Sinha, S., Bharadhwaj, H., Goyal, A., Larochelle, H., Garg, A., Shkurti, F.: Dibs: Diversity inducing information bottleneck in model ensembles. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 9666–9674 (2021)
38. Stickland, A.C., Murray, I.: Diverse ensembles improve calibration. In: *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning* (2020)
39. Tran, L., Veeling, B.S., Roth, K., Swiatkowski, J., Dillon, J.V., Snoek, J., Mandt, S., Salimans, T., Nowozin, S., Jenatton, R.: Hydra: Preserving ensemble diversity for model distillation. In: *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning* (2020)
40. Tsymbalov, E., Fedyanin, K., Panov, M.: Dropout strikes back: Improved uncertainty estimation via diversity sampling (2020)
41. Wen, Y., Tran, D., Ba, J.: Batchensemble: An alternative approach to efficient ensemble and lifelong learning. In: *Eighth International Conference on Learning Representations (ICLR 2020)* (2020)
42. Wenzel, F., Snoek, J., Tran, D., Jenatton, R.: Hyperparameter ensembles for robustness and uncertainty quantification. In: *Advances in Neural Information Processing Systems* (2020)
43. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. In: *Advances in neural information processing systems* (2020)
44. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., Schmidt, L.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time (2022). <https://doi.org/10.48550/ARXIV.2203.05482>
45. Zaidi, S., Zela, A., Elsken, T., Holmes, C., Hutter, F., Teh, Y.W.: Neural ensemble search for performant and calibrated predictions (2020)

## 6 Supplemental Material

### 6.1 Variants and ablation studies

Table 3 shows multiple variants of our *Sample Diversity* and the *ADP* regularizer. Increasing the batch size of **sampled images** to 512 (fourfold) increases the accuracy on the original and corrupted data even further, while also lowering the *ECE*, suggesting that there is more future potential in this approach. When *Sample Diversity* is combined with the *ADP* regularizer, we observe an improvement in all measured metrics suggesting that *Sample Diversity* combines constructively with other diversity regularizers. We suspect that this is due to the different behaviors of the regularizers (see Section 6.3) that emerge due to the different datasets the regularizers work on. Replacing the uniform noise with the tinyImageNet dataset reduces the overall accuracy on the original and corrupted data and increases the calibration error compared to our standard approach. We suspect the ever-new nature of newly sampled images to be more effective, than a limited pool of images, for diversity maximization on OOD data. Combining real datasets with augmentations and corruptions could be a future solution to the problem.

We replace the log-determinant regularization term in the *ADP* ( $ADP_{\chi^2}$ ) and *Sample Diversity* ( $SD_{\chi^2}$ ) regularizer, with the pair-wise  $\chi^2$  distance (see Eq. 4). This has the advantage that ensembles with more members than classes can be trained with diversity regularization, as otherwise for a matrix  $Y \in \mathbb{R}^{C \times M}$  ( $C - 1$  in case of ADP)

$$\det(Y^T \cdot Y) = 0, \text{ if } C < M \quad (6)$$

Table 3 shows that this formulation performs just as well while being numerically more stable, as it does not require a determinant and matrix-inversion operation and allows for arbitrary ensemble sizes. An ensemble of size 11 on CIFAR-10 in this case also benefits greatly from diversity regularization. This is a useful property for datasets with a low number of classes.

Finally, following recent work [17,26] we apply adversarial Fast-Gradient sign attacks [10] on the regularizer, to create more effective out-of-distribution images, on which to maximize the diversity. However, this approach only slightly increases the accuracy on the corrupted data (0.3%), while significantly increasing the computational time, suggesting that uniformly sampled images are already effective enough.

To test if regularization during test-time is necessary or if a functional diverse initialization is enough, we only apply the regularizers in a 3 epoch warm-up phase, then switching to standard cross-entropy training. We test only using *Sample Diversity* ( $\text{OrthoInit}_{OOD}$ ) and *Sample Diversity* combined with *ADP* ( $\text{OrthoInit}_{IID+OOD}$ ). These initialization do not lead to improvements, suggesting that constant regularization during training is necessary. This is an interesting avenue of research, as the diversity of the ensemble members seems to collapse to more similar solutions if no regularisation is applied.

**Table 3.** Comparison of different *Sample Diversity* (SD) and *ADP* variants on a TreeNet on CIFAR-10.

Reg.	Accuracy $\uparrow$		ECE $\downarrow$	
	org.	corr.	org.	corr.
ADP (base)	.917 $\pm$ .002	.535 $\pm$ .019	.024 $\pm$ .000	.180 $\pm$ .031
SD (base)	.916 $\pm$ .002	.545 $\pm$ .007	.040 $\pm$ .002	.213 $\pm$ .014
OrthoInit <sub>OOD</sub>	.917 $\pm$ .001	.516 $\pm$ .007	.036 $\pm$ .003	.232 $\pm$ .010
OrthoInit <sub>ID+OOD</sub>	.917 $\pm$ .002	.512 $\pm$ .006	.037 $\pm$ .001	.247 $\pm$ .004
SD (bs. 512)	.921 $\pm$ .001	.562 $\pm$ .012	.041 $\pm$ .001	.210 $\pm$ .010
SD+ADP	.919 $\pm$ .001	.560 $\pm$ .010	.027 $\pm$ .002	.150 $\pm$ .008
SD (adversarial)	.918 $\pm$ .000	.549 $\pm$ .013	.036 $\pm$ .006	.167 $\pm$ .026
SD (ImageNet)	.916 $\pm$ .004	.520 $\pm$ .007	.040 $\pm$ .002	.234 $\pm$ .014
SD $_{\chi^2}$	.919 $\pm$ .003	.545 $\pm$ .008	.041 $\pm$ .005	.222 $\pm$ .010
ADP $_{\chi^2}$	.921 $\pm$ .001	.533 $\pm$ .004	.027 $\pm$ .001	.208 $\pm$ .007
ind. (size 11)	.939 $\pm$ .001	.544 $\pm$ .005	.023 $\pm$ .001	.155 $\pm$ .003
ADP $_{\chi^2}$ (size 11)	.935 $\pm$ .004	.552 $\pm$ .007	.028 $\pm$ .001	.138 $\pm$ .007
SD $_{\chi^2}$ (size 11)	.932 $\pm$ 0.003	.589 $\pm$ .003	.027 $\pm$ .001	.114 $\pm$ .009

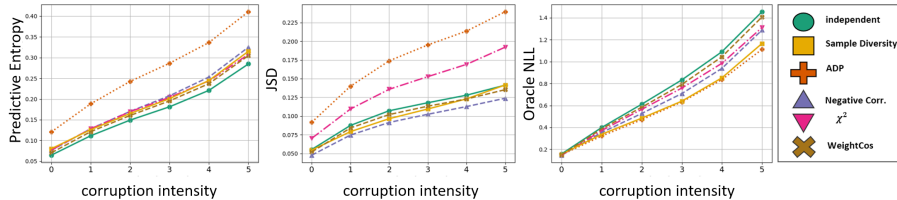
## 6.2 Distance in parameter space

We additionally use the distance in parameter space as a baseline regularizer to contrast distance in parameter space with functional diversity. Distance in parameter space is no guarantor of diverse functions, as different parameter settings can represent the same function, through a reparameterization of the function. A simple example of this is a permutation of filters inside a convolutional layer, with corresponding changes to the filters in adjacent layers. While this operation creates distance in parameter space, both parameter settings represent the same function.

Benjamin et al. [3] experimented with distance in function space, coming to the conclusion that parameter distance is no good measure for functional differences. We conduct similar experiments by introducing the *WeightCos* regularizer that orthogonalizes the parameter vectors  $\theta_i \in \mathbb{R}^P$ , stacked into a matrix  $\Theta \in \mathbb{R}^{M \times P}$ , of the individual ensemble members during training.

$$WeightCos(\theta_1, \dots, \theta_M) = \log(\det(\Theta \cdot \Theta^T))$$

We report the performance of the *WeightCos* regularizers in some of the later tables. While *WeightCos* does not lead to improvements in the Deep Ensemble and TreeNet architecture we observed that in the BatchEnsemble architecture *WeightCos* behaves very similar to the *ADP* regularizer in terms of *ECE* and *NLL*, leading to substantial improvements under dataset shift. We suspect that the formulation of the individual members in the BatchEnsemble does not allow for easy reparameterization. In this case orthogonality in parameter space could



**Fig. 5.** Entropy (left), Jensen-Shannon-Divergence (middle) and Oracle NLL (right) for a 5-member TreeNet on CIFAR-10, with different regularizations. The x-axis indicates the level of corruption.

induce functional diversity, which is an interesting finding for future work. See Section 6.7 for the results including the *WeightCos* regularizer.

### 6.3 Differences in diversity regularizers

To further investigate the effects diversity regularization has on the predictions of the ensemble, we measure the predictive entropy, Jensen-Shannon-Divergence, and Oracle NLL [27]. The predictive entropy of the ensemble is an indicator of how diverse an ensemble is, as more varying individual predictions will increase the entropy of the mean prediction.

$$H(\bar{y}) = -\frac{1}{\log(C)} \sum_{i=1}^C \bar{y}_i \log(\bar{y}_i) \quad (7)$$

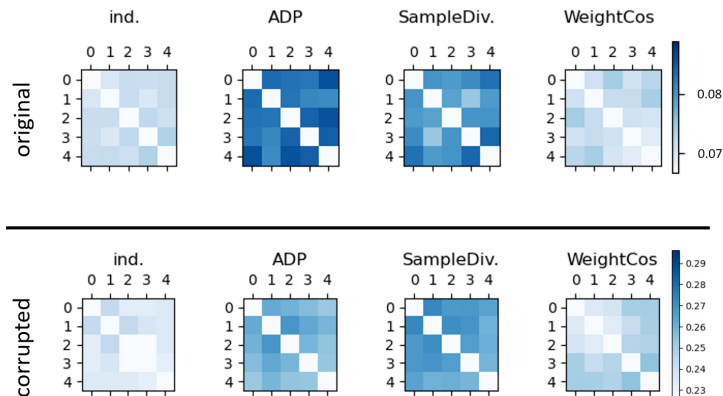
However, an ensemble composed of highly uncertain but similar members will also produce high entropy predictions. For this, we measure the Jensen-Shannon-Divergence (abbr.: JSD), which is the mean Kullback-Leibler (abbr.: KL) divergence between the individual ensemble members' predictions  $(y_1, \dots, y_M)$  and the mean prediction  $\bar{y}$ . A higher value for the *JSD* therefore indicates more diverse predictions across the ensemble.

$$KL(y_i || y_j) = -\sum_{k=1}^C y_i^{(k)} \cdot \log\left(\frac{y_i^{(k)}}{y_j^{(k)}}\right) \quad (8)$$

$$JSD(\bar{y}, y_1, \dots, y_M) = \frac{1}{M} \sum_{i=1}^M KL(y_i || \bar{y}) \quad (9)$$

A measure introduced by Lee et al. [27] is the Oracle NLL, which is the negative log-likelihood of the best performing ensemble member for each input. A more diversified ensemble with more specialized members results in a lower Oracle *NLL*.

The result for a TreeNet with 5 members can be found in Figure 5. The markers indicate the type of regularization and the x-axis indicates the level of



**Fig. 6.** Percentage of different argmax predictions between the ensemble members on the original data and the highest corruption level.

corruption. It can be seen that in the *ADP* regularized ensemble the individual members stray the furthest from the mean prediction, as seen in the higher entropy, Jensen-Shannon-Divergence between the members and also the significantly lower *Oracle NLL* [27]. On the other hand, the entropy of the *Sample Diversity* regularizer is only slightly increased compared to the independent ensemble training, even though the *ECE* and accuracy are constantly superior, which is most likely due to *Sample Diversity* only regularizing on out-of-distribution data, leaving the predictions on the training data intact. *Sample Diversity* and *ADP* have the lowest *Oracle NLL*, indicating a high functional diversity and specialised members. Consistent with our prior results, this is then followed by the  $\chi^2$  and *Negative Correlation* regularizer, which also performed worse on the other measured metrics. Independent training and *WeightCos* have the highest *Oracle NLL*, indicating that they lack diverse members. Additionally, *WeightCos* also has a very low *JSD*, showcasing that distance in parameter space is no guarantor of diverse members or diverse predictions. The *JSD* of the *Negative Correlation* regularizer is lower than that of the independent baseline, while the predictive entropy is higher. We interpret that as the *Negative Correlation* regularizer producing highly spread out and uncertain predictive distributions, which results in a lower *JSD*. These results show that all regularization approaches increase the differences between member predictions, as seen in the lower *Oracle NLL* and the higher entropy, but they do not all behave in the same way. We suspect these differences in the behaviour to be the reason why *ADP* and *Sample Diversity* combined so well in our experiments.

Figure 6 shows the percentage of differing argmax predictions on the original data and the highest corruption level. While distance in parameter space does not lead to more diverse predictions both *Sample Diversity* and *ADP* produce comparably diverse predictions.

## 6.4 Network capacity

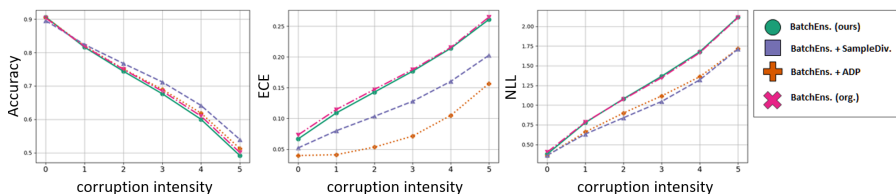
To test if diversity regularization also works with larger models we conduct experiments on the ResNet-44 architecture. Table 4 shows that when we use a bigger base architecture for the TreeNet, diversity training is still highly effective. In this case *ADP* even slightly increases the accuracy on the original data. Also in terms of *ECE* and *NLL* both regularizers lead to an improvement. However, further experiments with more varied architectures could give more insight in future work.

**Table 4.** Capacity experiments. To test if, diversity regularization still performs well with bigger architectures, the backbone of the TreeNet was exchanged for a ResNet-44 architecture (Res44). All experiments were conducted on CIFAR-10 with a TreeNet architecture with 5 members. We report the accuracy, *ECE* and *NLL*.

Method (corruption intensity)	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
	org.	corr.	org.	corr.	org.	corr.
ind.	.922 $\pm$ .001	.510 $\pm$ .016	.049 $\pm$ .005	.284 $\pm$ .004	.343 $\pm$ .031	2.175 $\pm$ .029
ADP	<b>.926</b> $\pm$ .002	.544 $\pm$ .004	<b>.032</b> $\pm$ .001	<b>.227</b> $\pm$ .031	<b>.318</b> $\pm$ .004	<b>1.853</b> $\pm$ .120
SampleDiv.	.922 $\pm$ .003	<b>.546</b> $\pm$ .008	.047 $\pm$ .002	.238 $\pm$ .021	.321 $\pm$ .025	1.887 $\pm$ .118

## 6.5 Differences in training Batch Ensemble

In our experiments, we trained the Batch Ensemble with the same data input in each step, while in the original paper [43] a batch was split over each member in every step. The difference is that in our approach each member sees the data points in the same order. This could lead to a reduction in diversity, which could be larger than the gain from diversity regularization. Figure 7 plots our *Batch Ensemble* training schedule (BatchEns. (ours)), the original training schedule (BatchEns. (org.)) and our schedule trained with *Sample Diversity* and *ADP*. While there is a minimal gain in the original training schedule, as noted by Ford et al. [7], the gain of the diversity regularization is far greater.



**Fig. 7.** Comparison of training the Batch Ensemble architecture with our training schedule compared to the original implementation.



## 6.6 Different split levels

We test using different split points, which we call split levels, in the TreeNet architecture to measure the influence of the ratio of shared parameters on the effectiveness of diversity regularization. In most of our experiments we set the split level to 3, which corresponds to a split just before the third ResNet block. A split of 0 is the same as a Deep Ensemble trained with the same data order during training. Split level 1 and 2 are placed before the first and second ResNet block, while split level 4 splits the network just before the last convolutional layer.

Table 5 shows our results for splitting a TreeNet trained on CIFAR-10 with 5 members at these different split points. As can be seen diversity regularization is effective even in the extreme case of split level 4, where just one convolutional layer is regularized.

**Table 5.** Experiments with different split levels on the TreeNet architecture on CIFAR-10 with 5 members. Comparison of the *Sample Diversity* and *ADP* regularizer.

Split Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$		
	org.	corr.	org.	corr.	org.	corr.	
0	ind.	<b>.936</b> $\pm$ .001	.543 $\pm$ .010	<b>.023</b> $\pm$ .001	.170 $\pm$ .014	<b>.210</b> $\pm$ .004	1.783 $\pm$ .051
	ADP	.933 $\pm$ .000	.549 $\pm$ .005	.032 $\pm$ .002	<b>.126</b> $\pm$ .010	.241 $\pm$ .004	<b>1.606</b> $\pm$ .037
	SampleDiv.	.933 $\pm$ .001	<b>.573</b> $\pm$ .004	.022 $\pm$ .001	.134 $\pm$ .002	.221 $\pm$ .004	1.634 $\pm$ .010
1	ind.	<b>.936</b> $\pm$ .001	.546 $\pm$ .014	<b>.024</b> $\pm$ .000	.163 $\pm$ .011	<b>.214</b> $\pm$ .004	1.769 $\pm$ .064
	ADP	.935 $\pm$ .001	.553 $\pm$ .005	.036 $\pm$ .004	<b>.108</b> $\pm$ .010	.245 $\pm$ .008	<b>1.545</b> $\pm$ .034
	SampleDiv.	.933 $\pm$ .002	<b>.570</b> $\pm$ .001	.024 $\pm$ .001	.148 $\pm$ .008	.214 $\pm$ .009	1.566 $\pm$ .029
2	ind.	<b>.932</b> $\pm$ .002	.530 $\pm$ .002	<b>.023</b> $\pm$ .001	.185 $\pm$ .009	<b>.226</b> $\pm$ .005	1.822 $\pm$ .063
	ADP	.931 $\pm$ .001	.539 $\pm$ .004	.028 $\pm$ .002	<b>.148</b> $\pm$ .014	.256 $\pm$ .003	<b>1.671</b> $\pm$ .046
	SampleDiv.	.929 $\pm$ .003	<b>.554</b> $\pm$ .009	.024 $\pm$ .001	.166 $\pm$ .016	.225 $\pm$ .007	1.624 $\pm$ .074
3	ind.	<b>.919</b> $\pm$ .001	.523 $\pm$ .007	.035 $\pm$ .001	.234 $\pm$ .012	<b>.286</b> $\pm$ .006	1.986 $\pm$ .076
	ADP	.917 $\pm$ .002	.535 $\pm$ .019	<b>.024</b> $\pm$ .000	<b>.180</b> $\pm$ .031	.298 $\pm$ .005	1.699 $\pm$ .132
	SampleDiv.	.916 $\pm$ 0.002	<b>.545</b> $\pm$ .007	.030 $\pm$ .002	.213 $\pm$ .014	.305 $\pm$ .013	1.822 $\pm$ .044
4	ind.	.899 $\pm$ .003	.502 $\pm$ .010	.067 $\pm$ .018	.257 $\pm$ .032	.395 $\pm$ .051	1.996 $\pm$ .209
	ADP	<b>.902</b> $\pm$ .001	.509 $\pm$ .024	<b>.038</b> $\pm$ .002	<b>.216</b> $\pm$ .027	<b>.393</b> $\pm$ .008	<b>1.843</b> $\pm$ .128
	SampleDiv.	.900 $\pm$ .001	<b>.527</b> $\pm$ .007	.081 $\pm$ .018	.243 $\pm$ .022	.438 $\pm$ .063	1.902 $\pm$ .093

### 6.7 Detailed results - CIFAR-10, CIFAR-100 and SVHN

Here we show the detailed results over all three datasets. Table 6 shows our results on CIFAR-10, Table 7 our results on the SVHN dataset and Table 8 our results on CIFAR-100. We notice that all regularizers have problems on CIFAR-100, which is most likely due to the large number of classes compared to the small number of ensemble members.

When we repeat the experiments with ensemble size 20 on CIFAR-100 we observe a better performance (see Table 9). However, datasets with large number of classes remain a problem. Here methods that utilize class number independent measures like internal activation’s of the neural network for diversification could prove superior.

### 6.8 Detailed results - Different Ensemble Sizes

Table 10 shows the results for our experiments on CIFAR-10 on the TreeNet architecture with different ensemble sizes, displaying the accuracy, *ECE* and *NLL*. We compare the ensemble sizes 2 to 5, using the *ADP* and *Sample Diversity* regularizer. Table 11 shows the same for the experiments on the BatchEnsemble and Deep Ensemble architectures. As before the scores are computed over 3 differently seeded runs and we report the mean and standard deviation. *Sample Diversity* consistently improves the accuracy under dataset shift and also lowers the *ECE* compared to the independent ensemble training (ind.). *ADP* performs best in terms of calibration error, having the lowest *ECE* in most settings. As mentioned before even a TreeNet of size 2 can outperform a Deep Ensemble of size 5 on the corrupted data, if diversity regularization is used. Furthermore, the  $ADP_{\chi^2}$  and  $SampleDiv.\chi^2$  formulation allow for training a Deep Ensemble of size 11 on CIFAR-10, which still provides large gains in terms of robustness to dataset shift.

**Table 6.** Experiments on CIFAR-10 with five members on different architectures.

Model	Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
		org.	corr.	org.	corr.	org.	corr.
DeepEns.	ind.	<b>.936</b> $\pm$ .001	.543 $\pm$ .010	<b>.023</b> $\pm$ .001	.170 $\pm$ .014	<b>.210</b> $\pm$ .004	1.783 $\pm$ .051
	ADP	.933 $\pm$ .000	.549 $\pm$ .005	.032 $\pm$ .002	<b>.126</b> $\pm$ .010	.241 $\pm$ .004	<b>1.606</b> $\pm$ .037
	NegCorr.	.934 $\pm$ .001	.538 $\pm$ .002	<b>.023</b> $\pm$ .001	.164 $\pm$ .007	<b>.210</b> $\pm$ .001	1.714 $\pm$ .082
	$\chi^2$	.934 $\pm$ .001	.542 $\pm$ .006	<b>.023</b> $\pm$ .000	.171 $\pm$ .008	.226 $\pm$ .007	1.767 $\pm$ .044
	SampleDiv.	.933 $\pm$ .001	<b>.579</b> $\pm$ .004	.022 $\pm$ .001	.134 $\pm$ .007	.217 $\pm$ .008	1.494 $\pm$ .026
	WeightCos.	.935 $\pm$ .001	.537 $\pm$ .007	<b>.023</b> $\pm$ .000	.166 $\pm$ .003	.212 $\pm$ .004	1.790 $\pm$ .034
TreeNet	ind.	.919 $\pm$ .002	.523 $\pm$ .01	.035 $\pm$ .001	.234 $\pm$ .010	.286 $\pm$ .006	1.990 $\pm$ .076
	ADP	.917 $\pm$ .002	.535 $\pm$ .019	<b>.024</b> $\pm$ .000	<b>.180</b> $\pm$ .031	.298 $\pm$ .005	1.699 $\pm$ .132
	NegCorr.	.918 $\pm$ .003	.528 $\pm$ .013	.027 $\pm$ .002	.200 $\pm$ .014	<b>.271</b> $\pm$ .009	1.785 $\pm$ .095
	$\chi^2$	<b>.920</b> $\pm$ .004	.517 $\pm$ .013	.027 $\pm$ .001	.238 $\pm$ .013	.282 $\pm$ .014	1.925 $\pm$ .117
	SampleDiv.	.916 $\pm$ .002	<b>.545</b> $\pm$ .007	.030 $\pm$ .002	.213 $\pm$ .014	.305 $\pm$ .013	1.822 $\pm$ .044
	WeightCos.	.919 $\pm$ .002	.517 $\pm$ .014	.032 $\pm$ .003	.225 $\pm$ .020	.275 $\pm$ .014	1.927 $\pm$ .143
BatchEns.	ind.	.905 $\pm$ .001	.512 $\pm$ .019	.097 $\pm$ .002	.285 $\pm$ .014	.455 $\pm$ .003	2.254 $\pm$ .099
	ADP	.906 $\pm$ .002	.517 $\pm$ .011	.032 $\pm$ .008	<b>.171</b> $\pm$ .049	.363 $\pm$ .036	1.735 $\pm$ .160
	NegCorr.	.904 $\pm$ .001	.503 $\pm$ .002	.072 $\pm$ .021	.258 $\pm$ .030	.385 $\pm$ .052	2.086 $\pm$ .230
	$\chi^2$	.905 $\pm$ .002	.503 $\pm$ .014	.058 $\pm$ .007	.265 $\pm$ .030	.391 $\pm$ .032	2.069 $\pm$ .178
	SampleDiv.	.904 $\pm$ .000	<b>.545</b> $\pm$ .007	.037 $\pm$ .015	.175 $\pm$ .032	<b>.343</b> $\pm$ .014	<b>1.649</b> $\pm$ .121
	WeightCos.	<b>.907</b> $\pm$ .003	.499 $\pm$ .005	<b>.022</b> $\pm$ .001	.182 $\pm$ .022	.385 $\pm$ .005	1.836 $\pm$ .108

**Table 7.** Experiments on SVHN with five members on different architectures.

Model	Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
		org.	corr.	org.	corr.	org.	corr.
TreeNet	ind.	.969 $\pm$ .001	.879 $\pm$ .000	<b>.008</b> $\pm$ .000	.026 $\pm$ .006	.126 $\pm$ .004	.432 $\pm$ .026
	ADP	<b>.971</b> $\pm$ .001	<b>.882</b> $\pm$ .004	.012 $\pm$ .001	.009 $\pm$ .003	.132 $\pm$ .005	.430 $\pm$ .014
	NegCorr.	.969 $\pm$ .001	.878 $\pm$ .004	.008 $\pm$ .001	.029 $\pm$ .009	.126 $\pm$ .004	.437 $\pm$ .042
	$\chi^2$	.969 $\pm$ .000	.880 $\pm$ .006	.009 $\pm$ .002	.027 $\pm$ .013	.128 $\pm$ .011	.434 $\pm$ .058
	SampleDiv.	.969 $\pm$ .001	<b>.882</b> $\pm$ .004	.009 $\pm$ .001	<b>.008</b> $\pm$ .001	<b>.122</b> $\pm$ .004	<b>.413</b> $\pm$ .018
BatchEns.	ind.	.965 $\pm$ .000	.877 $\pm$ .001	<b>.010</b> $\pm$ .001	.028 $\pm$ .008	.139 $\pm$ .001	.435 $\pm$ .008
	ADP	<b>.970</b> $\pm$ .001	<b>.891</b> $\pm$ .004	.021 $\pm$ .005	.036 $\pm$ .018	.139 $\pm$ .003	<b>.407</b> $\pm$ .007
	NegCorr.	.964 $\pm$ .003	.878 $\pm$ .007	<b>.010</b> $\pm$ .002	.025 $\pm$ .008	.138 $\pm$ .009	.432 $\pm$ .027
	$\chi^2$	.969 $\pm$ .002	.882 $\pm$ .006	<b>.010</b> $\pm$ .001	<b>.021</b> $\pm$ .020	<b>.131</b> $\pm$ .015	.417 $\pm$ .041
	SampleDiv.	.965 $\pm$ .000	.879 $\pm$ .002	<b>.010</b> $\pm$ .001	.029 $\pm$ .004	.139 $\pm$ .003	.442 $\pm$ .013
WeightCos.	.966 $\pm$ .001	.867 $\pm$ .003	.018 $\pm$ .004	.022 $\pm$ .011	.141 $\pm$ .003	.450 $\pm$ .007	

**Table 8.** Experiments on CIFAR-100 with five members on different architectures.

Model	Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
		org.	corr.	org.	corr.	org.	corr.
DeepEns.	ind.	<b>.726</b> $\pm$ .001	.300 $\pm$ .001	.055 $\pm$ .000	.058 $\pm$ .002	<b>1.008</b> $\pm$ .008	3.329 $\pm$ .008
	ADP	.719 $\pm$ .001	<b>.308</b> $\pm$ .002	.128 $\pm$ .004	<b>.032</b> $\pm$ .004	1.175 $\pm$ .003	3.274 $\pm$ .025
	NegCorr	.623 $\pm$ .015	.267 $\pm$ .002	.073 $\pm$ .003	.035 $\pm$ .002	1.375 $\pm$ .054	3.342 $\pm$ .016
	$\chi^2$	.717 $\pm$ .002	<b>.308</b> $\pm$ .002	.078 $\pm$ .002	.035 $\pm$ .002	1.020 $\pm$ .003	3.225 $\pm$ .019
	SampleDiv.	.708 $\pm$ .002	.306 $\pm$ .001	<b>.053</b> $\pm$ .003	.053 $\pm$ .003	1.070 $\pm$ .006	<b>3.173</b> $\pm$ .025
	WeightCos	.726 $\pm$ .003	.303 $\pm$ .002	.056 $\pm$ .001	.056 $\pm$ .002	<b>1.005</b> $\pm$ .003	3.287 $\pm$ .014
TreeNet	ind.	.710 $\pm$ .004	.288 $\pm$ .007	.036 $\pm$ .002	.075 $\pm$ .004	<b>1.054</b> $\pm$ .007	3.380 $\pm$ .038
	ADP	.708 $\pm$ .004	.292 $\pm$ .001	.106 $\pm$ .004	<b>.020</b> $\pm$ .002	1.207 $\pm$ .006	3.362 $\pm$ .025
	NegCorr.	.595 $\pm$ .003	.244 $\pm$ .008	.038 $\pm$ .003	.057 $\pm$ .008	1.451 $\pm$ .016	3.474 $\pm$ .077
	$\chi^2$	<b>.711</b> $\pm$ .003	.290 $\pm$ .001	.043 $\pm$ .002	.075 $\pm$ .007	1.076 $\pm$ .010	3.362 $\pm$ .019
	SampleDiv.	.701 $\pm$ .003	<b>.306</b> $\pm$ .004	<b>.034</b> $\pm$ .003	.082 $\pm$ .001	1.083 $\pm$ .011	<b>3.188</b> $\pm$ .036
	WeightCos	.705 $\pm$ .003	.283 $\pm$ .001	.038 $\pm$ .004	.076 $\pm$ .005	1.060 $\pm$ .005	3.437 $\pm$ .013
BatchEns.	ind.	.645 $\pm$ .000	.259 $\pm$ .005	.073 $\pm$ .005	.093 $\pm$ .011	1.336 $\pm$ .008	3.525 $\pm$ .027
	ADP	.637 $\pm$ .004	.265 $\pm$ .004	.082 $\pm$ .005	<b>.017</b> $\pm$ .003	1.537 $\pm$ .003	3.542 $\pm$ .055
	$\chi^2$	.642 $\pm$ .004	.258 $\pm$ .002	.055 $\pm$ .004	.067 $\pm$ .004	1.341 $\pm$ .019	3.484 $\pm$ .047
	SampleDiv.	.635 $\pm$ .003	<b>.275</b> $\pm$ .002	.073 $\pm$ .007	.102 $\pm$ .007	1.372 $\pm$ .007	<b>3.354</b> $\pm$ .002
	WeightCos	<b>.648</b> $\pm$ .002	.265 $\pm$ .003	<b>.038</b> $\pm$ .003	.075 $\pm$ .008	<b>1.277</b> $\pm$ .015	3.506 $\pm$ .026

**Table 9.** Experiments on CIFAR-100 with a TreeNet and ensemble size 20.

Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
	org.	corr.	org.	corr.	org.	corr.
ind.	<b>.721</b> $\pm$ .003	.297 $\pm$ .005	.043 $\pm$ .002	.083 $\pm$ .003	<b>.991</b> $\pm$ .009	3.345 $\pm$ .042
ADP	.718 $\pm$ .001	<b>.309</b> $\pm$ .006	.189 $\pm$ .002	<b>.048</b> $\pm$ .004	1.339 $\pm$ .013	3.319 $\pm$ .034
SampleDiv.	.719 $\pm$ .002	<b>.309</b> $\pm$ .007	<b>.042</b> $\pm$ .001	.087 $\pm$ .004	1.006 $\pm$ .008	<b>3.215</b> $\pm$ .052

**Table 10.** Experiments with different TreeNet ensemble sizes on CIFAR-10. Comparison of the *Sample Diversity* and *ADP* regularizer with independent training on different architectures under dataset shift.

Model	Size	Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
			org.	corr.	org.	corr.	org.	corr.
TreeNet	2	ind.	<b>.909</b> $\pm$ .002	.511 $\pm$ .005	.036 $\pm$ .000	.220 $\pm$ .013	<b>.310</b> $\pm$ .004	1.862 $\pm$ .026
		ADP	<b>.909</b> $\pm$ .000	.523 $\pm$ .012	<b>.028</b> $\pm$ .002	<b>.172</b> $\pm$ .015	.330 $\pm$ .007	1.696 $\pm$ .069
		SampleDiv.	.906 $\pm$ .002	<b>.541</b> $\pm$ .009	.042 $\pm$ .002	.192 $\pm$ .007	.332 $\pm$ .010	<b>1.680</b> $\pm$ .032
TreeNet	3	ind.	<b>.919</b> $\pm$ .002	.518 $\pm$ .008	.039 $\pm$ .002	.236 $\pm$ .003	<b>.295</b> $\pm$ .004	1.972 $\pm$ .012
		ADP	<b>.919</b> $\pm$ .002	.525 $\pm$ .006	<b>.025</b> $\pm$ .002	.189 $\pm$ .009	.301 $\pm$ .013	1.752 $\pm$ .029
		SampleDiv.	.910 $\pm$ .001	<b>.542</b> $\pm$ .014	.036 $\pm$ .001	<b>.187</b> $\pm$ .011	.303 $\pm$ .006	<b>1.693</b> $\pm$ .059
TreeNet	4	ind.	.918 $\pm$ .002	.515 $\pm$ .009	.034 $\pm$ .003	.226 $\pm$ .012	<b>.290</b> $\pm$ .008	1.933 $\pm$ .120
		ADP	<b>.919</b> $\pm$ .002	.524 $\pm$ .010	.026 $\pm$ .001	.190 $\pm$ .021	.297 $\pm$ .004	1.740 $\pm$ .101
		SampleDiv.	.910 $\pm$ .003	<b>.543</b> $\pm$ .014	.033 $\pm$ .001	<b>.185</b> $\pm$ .005	.298 $\pm$ .008	<b>1.654</b> $\pm$ .059
TreeNet	5	ind.	<b>.919</b> $\pm$ .002	.523 $\pm$ .01	.035 $\pm$ .001	.234 $\pm$ .010	<b>.286</b> $\pm$ .006	1.990 $\pm$ .076
		ADP	.917 $\pm$ .002	.535 $\pm$ .019	<b>.024</b> $\pm$ .000	<b>.180</b> $\pm$ .031	.298 $\pm$ .005	1.699 $\pm$ .132
		SampleDiv.	.916 $\pm$ .002	<b>.545</b> $\pm$ .007	.030 $\pm$ .001	.213 $\pm$ .012	.290 $\pm$ .005	<b>1.659</b> $\pm$ .062

**Table 11.** Experiments with different Batch Ensemble and Deep Ensemble ensemble sizes on CIFAR-10. Comparison of the *Sample Diversity* and *ADP* regularizer with independent training on different architectures under dataset shift.

Model	Size Method	Accuracy $\uparrow$		ECE $\downarrow$		NLL $\downarrow$	
		org.	corr.	org.	corr.	org.	corr.
BatchEns. 2	ind.	<b>.898</b> $\pm$ .005	.499 $\pm$ .010	.046 $\pm$ .016	.220 $\pm$ .048	.348 $\pm$ .027	1.883 $\pm$ .109
	ADP	.893 $\pm$ .001	.518 $\pm$ .007	<b>.032</b> $\pm$ .004	<b>.118</b> $\pm$ .002	.367 $\pm$ .007	<b>1.596</b> $\pm$ .035
	SampleDiv.	.897 $\pm$ .001	<b>.533</b> $\pm$ .006	.038 $\pm$ .005	.186 $\pm$ .008	<b>.337</b> $\pm$ .011	1.711 $\pm$ .036
BatchEns. 3	ind.	.905 $\pm$ .002	.503 $\pm$ .013	.085 $\pm$ .015	.270 $\pm$ .019	.430 $\pm$ .038	2.159 $\pm$ .112
	ADP	<b>.906</b> $\pm$ .002	.516 $\pm$ .010	<b>.035</b> $\pm$ .003	<b>.135</b> $\pm$ .015	<b>.335</b> $\pm$ .005	<b>1.678</b> $\pm$ .060
	SampleDiv.	.897 $\pm$ .001	<b>.534</b> $\pm$ .002	.039 $\pm$ .018	.187 $\pm$ .028	.344 $\pm$ .027	1.754 $\pm$ .084
BatchEns. 4	ind.	<b>.906</b> $\pm$ .001	.491 $\pm$ .001	.067 $\pm$ .018	.261 $\pm$ .038	.377 $\pm$ .054	2.117 $\pm$ .170
	ADP	.905 $\pm$ .001	.513 $\pm$ .002	<b>.040</b> $\pm$ .011	<b>.157</b> $\pm$ .055	<b>.360</b> $\pm$ .018	1.716 $\pm$ .085
	SampleDiv.	.896 $\pm$ .002	<b>.540</b> $\pm$ .011	.052 $\pm$ .010	.203 $\pm$ .016	.363 $\pm$ .020	<b>1.712</b> $\pm$ .096
BatchEns. 5	ind.	.905 $\pm$ .001	.512 $\pm$ .019	.097 $\pm$ .002	.285 $\pm$ .014	.455 $\pm$ .003	2.254 $\pm$ .099
	ADP	<b>.906</b> $\pm$ .002	.517 $\pm$ .011	<b>.032</b> $\pm$ .008	<b>.171</b> $\pm$ .049	.363 $\pm$ .036	1.735 $\pm$ .160
	SampleDiv.	.904 $\pm$ .000	<b>.545</b> $\pm$ .007	.037 $\pm$ .015	.175 $\pm$ .032	<b>.343</b> $\pm$ .014	<b>1.649</b> $\pm$ .121
DeepEns. 2	ind.	.921 $\pm$ .004	.519 $\pm$ .009	<b>.022</b> $\pm$ .001	.191 $\pm$ .009	<b>.248</b> $\pm$ .010	1.786 $\pm$ .039
	ADP	<b>.922</b> $\pm$ .002	.541 $\pm$ .002	.027 $\pm$ .001	<b>.143</b> $\pm$ .007	.282 $\pm$ .002	<b>1.599</b> $\pm$ .010
	SampleDiv.	.921 $\pm$ .002	<b>.569</b> $\pm$ .012	.029 $\pm$ .001	.158 $\pm$ .009	.263 $\pm$ .005	1.579 $\pm$ .062
DeepEns. 3	ind.	<b>.929</b> $\pm$ .002	.532 $\pm$ .009	<b>.021</b> $\pm$ .001	.174 $\pm$ .017	<b>.225</b> $\pm$ .005	1.747 $\pm$ .058
	ADP	.928 $\pm$ .001	.544 $\pm$ .002	.029 $\pm$ .003	<b>.132</b> $\pm$ .008	.259 $\pm$ .010	1.592 $\pm$ .035
	SampleDiv.	.926 $\pm$ .001	<b>.566</b> $\pm$ .014	.026 $\pm$ .002	.149 $\pm$ .021	.238 $\pm$ .002	<b>1.562</b> $\pm$ .066
DeepEns. 4	ind.	<b>.932</b> $\pm$ .004	.539 $\pm$ .005	<b>.023</b> $\pm$ .001	.157 $\pm$ .006	<b>.216</b> $\pm$ .009	1.675 $\pm$ .038
	ADP	.929 $\pm$ .001	.554 $\pm$ .002	.031 $\pm$ .002	<b>.113</b> $\pm$ .009	.250 $\pm$ .008	1.532 $\pm$ .006
	SampleDiv.	.930 $\pm$ .002	<b>.575</b> $\pm$ .003	<b>.023</b> $\pm$ .001	.129 $\pm$ .007	.223 $\pm$ .002	<b>1.499</b> $\pm$ .010
DeepEns. 5	ind.	<b>.936</b> $\pm$ .001	.543 $\pm$ .010	.023 $\pm$ .001	.170 $\pm$ .014	<b>.210</b> $\pm$ .004	1.783 $\pm$ .051
	ADP	.933 $\pm$ .000	.549 $\pm$ .005	.032 $\pm$ .002	<b>.126</b> $\pm$ .010	.241 $\pm$ .004	1.606 $\pm$ .037
	SampleDiv.	.933 $\pm$ .001	<b>.579</b> $\pm$ .004	<b>.022</b> $\pm$ .001	.134 $\pm$ .007	.217 $\pm$ .008	<b>1.494</b> $\pm$ .026
DeepEns. 11	ind.	<b>.939</b> $\pm$ .001	.544 $\pm$ .005	<b>.023</b> $\pm$ .001	.155 $\pm$ .003	<b>.191</b> $\pm$ .001	1.667 $\pm$ .043
	ADP $_{\chi^2}$	.935 $\pm$ .004	.552 $\pm$ .007	.028 $\pm$ .001	.138 $\pm$ .007	.206 $\pm$ .010	1.524 $\pm$ .030
	SampleDiv. $_{\chi^2}$	.932 $\pm$ .003	<b>.589</b> $\pm$ .003	.027 $\pm$ .001	<b>.114</b> $\pm$ .009	.208 $\pm$ .009	<b>1.420</b> $\pm$ .034

---

## Contribution and impact

As long as the models within an ensemble are similar, a larger number of members does not translate into better uncertainty or performance scores. Instead, only the computational overhead is increased. This has spurred research into increasing ensemble diversity, hoping to obtain smaller yet more effective ensembles.

While some methods add variety in terms of model architecture or data augmentations, *functional diversity* approaches directly encourage diversity in the solution space by regularizing models to produce different predictions. This strategy can augment any architecture and does not require making additional design choices. Nevertheless, the competing objective can harm performance: by forcing some models to make wrong predictions, we may reduce the accuracy of the ensemble.

In this work, we introduce the *Sample Diversity* method that improves calibration and robustness under dataset shift by *only encouraging a diverse solution space for OOD samples*. By increasing diversity only for samples for which the model would anyhow be less confident, we can identify OOD samples more easily during testing without affecting the performance on ID data.

We demonstrate that our approach is particularly effective for weight-sharing architectures – specifically *TreeNets* (Lee et al., 2015) and *Batch Ensembles* (Wen et al., 2020) – that have a smaller computational footprint but typically less diversity.

## Discussion and limitations

Though the results in the paper are encouraging, most of our experiments were carried out on low-resolution computer vision datasets. This was due to the computational overhead involved with training multiple ensembles in many different settings. By staying in the domain of natural images, we could also use the widely popular corruptions introduced by Hendrycks and Dietterich (2018) in our evaluation.

In the preliminary experiments we carried out for medical image segmentation, we did not see an improvement in our method when compared to the existing *ADP* (Pang et al., 2019) approach that increases diversity in the in-distribution training data. This was potentially due to our overly simplistic strategy to generate OOD cases by sampling from a uniform distribution, which would naturally not be suitable for complex tasks. In future work, I would like to explore different ways to generate OOD cases that are semantically meaningful for the task at hand.

Of course, an additional downside of the proposed method is that it forces the user to consider uncertainty estimation from the time the model is trained. The additional term in the loss function could also make training more difficult in some cases.

---

## 4.2. Conclusions and outlook

---

Within uncertainty estimation, training ensembles is a very simple strategy that identifies high-uncertainty cases for in-distribution data. Weight-sharing architectures have been proposed to limit the computational overhead of training and performing inference with different models (Lee et al., 2015; Wen et al., 2020), and by purposely increasing diversity among members we can obtain the advantages of increased performance and uncertainty estimation with only a few models.

---

## 5. Assessing the Coherence of Model Predictions

---

For some problems – such as classification or regression – simply observing the network output gives little information on its quality, so silent failures can go undetected. Fortunately, for tasks such as semantic segmentation and registration, we expect certain coherence in the obtained predictions. Just like an expert observer may immediately detect an incorrect mask that segments the hippocampus outside the brain, we may perform simple automatic tests that help us catch low-quality predictions without manual examination by an expert.

This has several implications. First, we can ensure that only coherent predictions reach the clinicians, avoiding the risk of them losing trust in the ML system. Secondly, we can control the quality of annotations that will serve as training data for a continuously adapting system, which may have been fully manually delineated or initialized by another ML system. Finally, we may employ such methods to get an idea of how a system deployed at different sites is performing.

What method we employ is highly dependent on the problem and whether we have access to the image data. Existing strategies for assessing the quality of segmentation masks typically train a model that predicts the quality (Valindria et al., 2017; Chen et al., 2020; Lee et al., 2020) or leverage domain knowledge to design meaningful features based on the expected shape (van Rikxoort et al., 2009). The second strategy has the advantage that the user can easily interpret why a certain prediction was discarded. In the following, we show how such an approach can be meaningful even when our anatomy has no defined number of connected components or a specific geometric shape.

---

### 5.1. The paper: Quality monitoring of federated Covid-19 lesion segmentation

---

While OOD detection methods observe the network inputs or activations and uncertainty estimation approaches look at outputs, here we only work with the predictions. We conceived the work that we published as *Quality monitoring of federated Covid-19 lesion segmentation* (González et al., 2022b) as preparation for a multi-clinical federated learning study for the segmentation of Covid-19-related findings in chest CT. The challenge lay in developing methods that would give a user in a central location an overview of how well a federated segmentation model was performing at each site, without ever having access to patient data.

We initially presented our work at the 102<sup>nd</sup> *German Röntgen Congress (RöKo)* in Remscheid on November 5<sup>th</sup>, 2021. The paper was accepted for oral presentation at the *Bildverarbeitung für die Medizin (BVM)* in June 27<sup>th</sup>, 2022, in Heidelberg, where it was nominated for the *Best Scientific Work* award.



# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Quality monitoring of federated Covid-19 lesion segmentation**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Quality monitoring of federated Covid-19 lesion segmentation" (González et al. 2022) was published as a full research paper at the "Bildverarbeitung für die Medizin (BVM)". It constitutes a joint work of Camila González, Christian Harder, Amin Ranem, Ricarda Fischbach, Isabel Kaltenborn, Armin Dadras, Andreas Bucher and Anirban Mukhopadhyay.*

*This work was supported by the RACOON network under BMBF, grant number [01KX2021].*

*As corresponding and leading author, C. González led the overall research design, management and writing process of the paper. C. Harder contributed the literature review. The choice of methodological framework and experimental framework were done by C. González and C. Harder together. A. Bucher, R. Fischbach and I. Kaltenborn collected the in-house data and reviewed the manuscript from a clinical perspective. C. Harder and A. Ranem prepared and pre-processed the openly available data and implemented the code. The methodology, results and discussion were written by C. González and C. Harder. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor of this work, who also contributed with continuous feedback during all phases of the paper writing process. All authors agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum: 01 / 09 / 2023 01 / 10 / 2023 01 / 10 / 2023 01 / 09 / 2023

Unterschrift:    

Camila González Christian Harder Amin Ranem Ricarda Fischbach

Datum: 01 / 12 / 2023 01 / 10 / 2023 01 / 10 / 2023 01 / 12 / 2023

Unterschrift:    

Isabel Kaltenborn Armin Dadras Andreas Bucher Anirban Mukhopadhyay

# Quality monitoring of federated Covid-19 lesion segmentation

Camila González<sup>1</sup>, Christian L. Harder<sup>1</sup>, Amin Ranem<sup>1</sup>, Ricarda Fischbach<sup>2</sup>,  
Isabel J. Kaltenborn<sup>2</sup>, Armin Dadras<sup>2</sup>, Andreas M. Bucher<sup>2</sup>, Anirban  
Mukhopadhyay<sup>1</sup>

<sup>1</sup> Medical and Environmental Computing, Technische Universität Darmstadt

<sup>2</sup> Diagnostische und Interventionelle Radiologie, Universitätsklinikum Frankfurt  
camila.gonzalez@gris.tu-darmstadt.de

**Abstract.** Federated Learning is the most promising way to train robust Deep Learning models for the segmentation of Covid-19-related findings in chest CTs. By learning in a decentralized fashion, heterogeneous data can be leveraged from a variety of sources and acquisition protocols whilst ensuring patient privacy. It is, however, crucial to continuously monitor the performance of the model. Yet when it comes to the segmentation of diffuse lung lesions, a quick visual inspection is not enough to assess the quality, and thorough monitoring of all network outputs by expert radiologists is not feasible. In this work, we present an array of lightweight metrics that can be calculated locally in each hospital and then aggregated for central monitoring of a federated system. Our linear model detects over 70% of low-quality segmentations on an out-of-distribution dataset and thus reliably signals a decline in model performance.

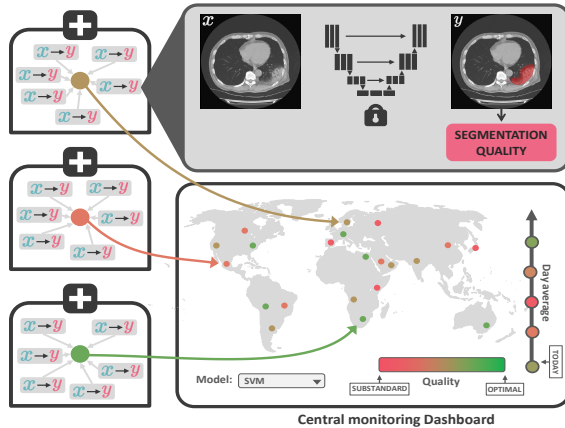
## 1 Introduction

The Covid-19 pandemic has strained medical resources across the world while demonstrating the value of time-saving workflow enhancements. Deep Learning solutions for the quantification of clinically relevant infection parameters, which segment Covid-19-characteristic lesions in CTs, have shown promising results.

Yet sufficient maturity for clinical use is frequently not reached by present approaches [1]. This is mainly due to neural networks failing silently coupled with a lack of appropriate quality controls. Scanner models and acquisition protocols vary between and within hospitals, changing image distribution. This causes deep learning models to produce low-quality outputs with high confidence [2].

Covid-19-related ground glass opacities and consolidations can occur in various forms, from covering multiple small regions to diffuse affection of the entire lung [3]. Identifying low-quality segmentation masks is very time consuming and requires extensive experience, but thorough monitoring of all network outputs by expert readers is not logistically feasible.

**Fig. 1.** Quality features are extracted and an SVM model is used to perform inference locally at several hospitals. These quality scores are aggregated for each site and visualized at a central dashboard. In the entire process, only the privacy-preserving aggregated scores leave the institutions.



Automated quality assurance for segmentation masks is not yet a developed field. Existing approaches include the training of a CNN on the logits of the segmentation prediction [4] or the concept of a Reverse Classification Algorithm [5] to predict segmentation quality. These are either computationally expensive or depend on rigid target shapes, which is not given in the case of Covid-19 lesions. Failed segmentations can however be identified by observing certain properties in the segmentation masks.

We propose an array of **lightweight yet reliable quality metrics for segmentation masks that do not require ground truth annotations**. These can be **calculated locally without the need for expert reader review and then aggregated for each hospital for central monitoring of federated systems**, as illustrated in Fig. 1.

## 2 Materials and Methods

We implemented our code with Python 3.8 and PyTorch 1.6 and performed a retrospective study using several open-source datasets, as well as in-house data. The code can be found at [github.com/MECLabTUDA/QA\\_Seg](https://github.com/MECLabTUDA/QA_Seg).

**Data:** To obtain a dataset of predicted segmentations, we extracted predictions from an nnU-Net [6] trained on the COVID-19 Lung Lesion Segmentation Challenge (*Challenge*) dataset [7]. We also predicted segmentations on *MosMed* [8], as well as in-house data with further 50 cases. Images were interpolated to dimension (50,512,512). Further details can be found in Table 1. We partitioned the predictions into in-distribution (ID) for the Challenge and in-house datasets (with which we trained our classifiers) and out-of-distribution (OOD) for MosMed. The ID datasets were randomly divided into *ID train* and *ID test*. We considered the Dice between ground truth and predicted masks as a measure of segmentation quality, as it is the most-used metric for segmentation overlap.

**Table 1.** Data distribution, including ratio of infection within the segmented lung volume [9], nnU-Net performance and number of failed segmentation masks.

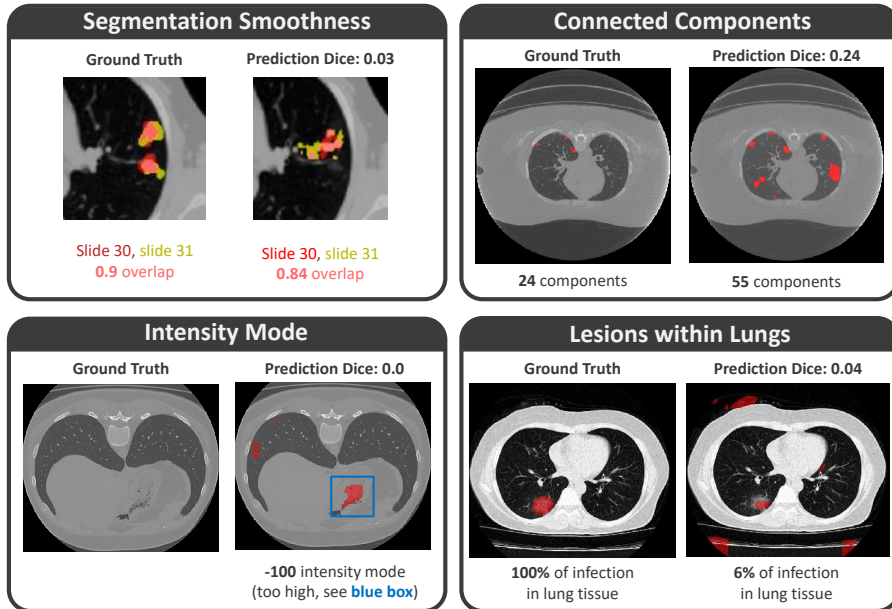
Property	Challenge	In-house	MosMed
<b>Nr. cases (train, test)</b>	199 (160, 39)	50 (40, 10)	50 (0, 50)
<b>Mean resolution</b>	(68.87,512.0,512.0)	(266.64,819.20,825.68)	(40.98,512.00,512.00)
<b>Infection ratio</b>	$0.061 \pm 0.093$	$0.275 \pm 0.274$	$0.016 \pm 0.015$
<b>nnU-Net Dice (train)</b>	$0.75 \pm 0.14$	$0.59 \pm 0.2$	N.A.
<b>nnU-Net Dice (test)</b>	$0.71 \pm 0.18$	$0.68 \pm 0.1$	$0.47 \pm 0.19$
<b>Failed masks (train)</b>	24	12	N.A.
<b>Failed masks (test)</b>	8	1	37

As shown in Table 1, the ID data is heavily skewed towards good-quality segmentations. We define a *failed* segmentation as having a Dice lower than 0.6 (following Valindria et al. [5]) and report their prevalence in Table 1.

**Proposed features:** Inspired by van Rikxoort et al. [10], we looked to predict the quality of segmentation masks - in the form of Dice coefficient - using only four features (see Fig. 2), defined as follows:

- *Connected Components:* While lung lesions may occupy several components, failed segmentations are often more disconnected. We counted the number of connected components using Scikit-Image [11], defining a component as one with a maximal distance of 3 by the City Block Metric to other voxels.
- *Intensity Mode:* Observing the intensity values in the CT, we can identify tissue that is very unlikely to be infected. Inspired by Kalka et al. [12], we fitted a Gaussian distribution over the largest component and returned its mean.
- *Segmentation Smoothness:* In a correct segmentation mask, we expect two consecutive slices to have a high overlap and thus a high two-dimensional Dice. We computed the smoothness for every component by taking the average Dice scores for all consecutive slices that were not identical. We then averaged the smoothness over all components.
- *Lesions within Lungs:* A correct segmentation mask should be completely contained within the lung. To factor this in, we used a pre-trained lung segmentation model [9] and recorded the percentage of segmented tissue that is inside of the lung.

**Models and training:** With these features, we trained and evaluated several models to predict the segmentation quality. We directly regressed the quality with a Ridge Regression (RR) and a Support Vector Regression (SVR) (trained until convergence) as well as a Multi-Layer-Perceptron (MLP) with (50,100,100,50) layers for 200 epochs minimizing the Mean Squared Error. We also discretized the quality values into five bins and performed classification with Support Vector Machine (SVM) and Logistic Regression (LR) models, using balanced class weights. Unless otherwise stated, we used the default Scikit-learn [13] library implementations.



**Fig. 2.** Exemplary subjects and slides for the four features used to assess segmentation quality.

**Evaluation:** As we were primarily interested in detecting failed segmentations, we report the sensitivity of all 5 models on this task. We also report the specificities for identifying the correct quality interval (averaged over 5 bins) on all ID and OOD datasets. In addition, we report the Mean Absolute Error as a metric that quantifies the ability of all models to directly predict the segmentation quality.

### 3 Results

In terms of sensitivity (detection of faulty segmentations) the classifiers (LR and SVM) outperformed the regression models by a large margin (see Table 2). This can be attributed to the class weights of the LR and SVM models balancing the disparately appearing classes in the training data, which improved their performance on differently distributed data. Though we were unable to detect the single failed segmentation out of 10 on the in-house dataset, we highlight the performance of the LR model, which detects over 60% of failed segmentations on both of the bigger Challenge and MosMed datasets. All models showed a high specificity of over 0.8 on all datasets. The regression models achieved a lower mean absolute error but seemed to overfit the good-quality segmentations on the training dataset, which might explain their worse sensitivity.

We further evaluated the LR model using 10000 bootstrapping runs, sampling 192 data points from the training set and evaluating the model’s sensitivity

**Table 2.** Sensitivity of finding failed segmentations (Dice < 0.6), specificity of identifying the correct quality interval (avg. over 5 bins) and Mean Absolute Error (mean+/- std) results for each model for ID and OOD datasets.

	Classifiers			Regressors		
	LR	SVM	RR	SVR	MLP	
Sensitivity	Challenge	0.63 (5/8)	0.38 (3/8)	0.38 (3/8)	0.13 (1/8)	0.25 (2/8)
	In-house	0.0 (0/1)	0.0 (0/1)	0.0 (0/1)	0.0 (0/1)	0.0 (0/1)
	MosMed	0.76 (28/37)	0.68 (25/37)	0.14 (5/37)	0.35 (13/37)	0.35 (13/37)
Specificity	Challenge	0.84	0.85	0.88	0.87	0.87
	In-house	0.8	0.83	0.95	0.9	0.9
	MosMed	0.8	0.83	0.82	0.84	0.85
MAE	Challenge	0.29 ± 0.22	0.26 ± 0.22	0.1 ± 0.1	0.11 ± 0.11	0.18 ± 0.13
	In-house	0.24 ± 0.12	0.26 ± 0.14	0.08 ± 0.09	0.1 ± 0.07	0.09 ± 0.07
	MosMed	0.33 ± 0.19	0.29 ± 0.23	0.22 ± 0.16	0.21 ± 0.18	0.23 ± 0.18

trained on these samples on the ID and OOD datasets for every run. We achieved 95% confidence intervals for the sensitivity covering a range from 0.22 to 1.0. Furthermore, using a p-valued test with a significance level of 0.05, we can reject every null hypothesis stating that the sensitivity of the LR model is below 0.28.

In order to evaluate the individual contribution of each feature, we performed an ablation study where we left out each of the features for LR models. The "Intensity Mode" feature proved to be the least useful. Leaving it out allows us to correctly identify 5 more high-quality segmentations as such, though 9 faulty segmentations less are detected. All in all, using all four features achieves the best sensitivity-to-specificity trade-off.

We attribute most of the falsely classified segmentations to the low representation of bad segmentations in the training data and to these displaying plausible shapes. For example, segmentation masks covering only a few spots of healthy lung tissue, containing intensity values of possibly infected areas, while maintaining a smooth shape, were not detected.

## 4 Discussion

We introduced a simple method to monitor performance of an nnU-Net trained to detect lung infections onset by Covid-19. We designed four features and found that a LR model using these reliably detects faulty segmentation masks. All the features are lightweight and do not require ground truth annotations, and so they can be used to monitor the deployment of a distributed, federated learning system.

Our findings have some limitations. First, we tested our methods retrospectively on a statically trained nnU-Net. This allowed us to accurately evaluate our methods, as we had access to ground truth test annotations, but a prospective

study on a federated system with a few participating institutions would better emulate real deployment.

Secondly, the CT data was acquired on ICU patients, thus introducing considerable bias in patient demographics which are likely not representative of the general Covid-19 population. This also suggests that a measure other than Dice may be better suited for the general population, as the expressiveness of Dice is heavily dependent on lesion size.

Finally, each dataset was annotated by a different group of experts, so the definitions of the findings may vary across datasets. This is often the case when evaluating with OOD data but should be taken into account when considering the differences in performance.

In conclusion, training models in a federated fashion allows to leverage heterogeneous data sources without compromising patient privacy. However, it is necessary to constantly monitor the quality of the model outputs. In this work, we introduced an array of **lightweight quality metrics that can be calculated locally and aggregated for central monitoring**. These are particularly well-suited to the use case of lung lesion segmentation in chest CTs, as lesions vary greatly in terms of form and location and verifying their correctness is time-intensive even for trained radiologists. Future work should expand the metric catalogue and assess the effectiveness of the proposed methods in a model deployed across multiple hospitals. Our results present a first step towards an effective quality control of federated lung lesion segmentation.

## References

1. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*. 2021;3(3):199–217.
2. Gonzalez C, Gotkowski K, Bucher A, et al. Detecting when pre-trained nnU-Net models fail silently for Covid-19 lung lesion segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2021. p. 304–314.
3. Bai HX, Hsieh B, Xiong Z, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology*. 2020;296(2):E46–E54.
4. Chen X, Men K, Chen B, et al. CNN-based quality assurance for automatic segmentation of breast cancer in radiotherapy. *Frontiers in Oncology*. 2020;10:524.
5. Valindria VV, Lavdas I, Bai W, et al. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*. 2017;36(8):1597–1606.
6. Isensee F, Jaeger PF, Kohl SA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*. 2021;18(2):203–211.
7. Roth H, Xu Z, Diez CT, et al. Rapid artificial intelligence solutions in a pandemic - the COVID-19-20 lung CT lesion segmentation challenge. *Research Square*. 2020;.
8. Morozov S, Andreychenko A, Pavlov N, et al. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:200506465*. 2020;.

9. Hofmanninger J, Prayer F, Pan J, et al. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*. 2020;4(1):1–13.
10. van Rikxoort EM, de Hoop B, Viergever MA, et al. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical physics*. 2009;36(7):2934–2947.
11. Van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453.
12. Kalka N, Bartlow N, Cukic B. An automated method for predicting iris segmentation failures. In: 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems. IEEE; 2009. p. 1–8.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *the Journal of machine Learning research*. 2011;12:2825–2830.



---

## Contribution and impact

As touched upon in Section 3.1.2, assessing the validity of lung lesion segmentation masks can be complex even for trained radiologists. This is due to their diffuse shapes and the fact that the intensities they take in the Hounsfield scale overlap with that of common CT artifacts. Intuitively, this does not seem like a problem that we can solve by performing simple quality checks on the predictions.

In this paper, we show that we can indeed identify many failure cases by only calculating four simple features: (1) the continuity of the two-dimensional masks across slices, (2) the number of connected components, (3) the intensity mode of the voxels covered by the segmentation and (4) ensuring that the prediction is within the lungs, for which we employ a separate lung segmentation model that is highly accurate (Hofmanninger et al., 2020).

This has a significant impact on model monitoring, as it means we can detect a significant performance degradation without any access to patient images or DNN activations (which could potentially encode patient-identifying information). We show in the paper how we can train various models with the designed features to get a continuous quality score for each patient. However, the fact that these are highly interpretable provides valuable insights into what is problematic about a particular prediction.

Considering that calculating the proposed features requires minimal hardware, this strategy can be employed in scenarios where we do not even have direct access to the network outputs but rather only to aggregated quality metrics, as the metrics could be calculated at the individual sites.

## Discussion and limitations

The study has many limitations, primarily with regard to feature design. The feature catalog could be tuned further, and new features could be incorporated. Additionally, a prospective study where the usability of the metrics is assessed in a real setting would be interesting, and this is hopefully an analysis we can perform in the future across the network of German university hospitals. It would also be interesting to assess whether failure cases are identified that OOD detection or uncertainty estimation methods fail to uncover.

Nevertheless, the study highlights how even state-of-the-art models can produce predictions that are clearly faulty. Many of these can be detected with simple, computationally inexpensive methods. Therefore, for problems such as semantic segmentation, the usability of metrics that directly assess the coherence of a prediction should not be underestimated.

---

## 5.2. Conclusions and outlook

---

Working with medical images introduces certain challenges, such as needing to work with small datasets made of high-dimensional data. But the importance of domain knowledge can help reduce the complexity of the problem. Particularly for quality assurance, automatic methods that consider such domain information can be more helpful than purely algorithmic techniques for detecting which model predictions can be trusted.

---

## 6. Domain Adaptation and OOD Generalization

---

Faced with the problem of domain shift, one may first attempt to avoid learning spurious correlations and idiosyncrasies in the data. The hope is that by encouraging the model to instead learn semantically meaningful representations, it will generalize better and be more robust in the presence of domain shift. Such an approach is especially possible if we have training data - with or without corresponding ground truth annotations - from several sources at once during training. Then, we can directly steer the training process so that the performance is stable across domains. We commonly refer to this as *out-of-distribution (OOD) generalization* (Hendrycks et al., 2021).

Another alternative, if we have a model trained for one particular *source* domain and wish to extract a prediction for an image from a different *target* domain, is to employ techniques from *domain adaptation* (Farahani et al., 2021), which allow us to align the latent space of the new image to our source domain.

These are exciting fields of research and highly relevant for developing models that perform as expected in the open world. During my doctoral studies, I had the pleasure of exploring these possibilities for medical image segmentation alongside several students, and I will briefly summarize our findings in the following sections. Our empirical results show that we *can* learn more generalizable features that are relevant for the downstream diagnosis. However, this often comes at the cost of more computationally expensive and challenging training procedures, where several hyperparameters need to be properly tuned in order to reach convergence.

---

### 6.1. Training models to maintain stable performance across domains

---

Some methods adapt the training objective to minimize performance variations between domains. Popular examples of taking this approach are *Invariant Risk Minimization (IRM)* (Peters et al., 2016) and *Variance Risk Extrapolation (VREx)* (Krueger et al., 2021). Of course, these methods assume that there is training data available from several domains during training. *VREx* in particular recognizes that the data shifts observed in the real world may be different – and likely more pronounced – than in the training corpus. Still, by minimizing the risk on the worse-performing domain, the resulting model can produce higher-quality results on yet-unseen domains.

Another strategy consists of minimizing the ability of a *domain predictor* to identify the domain of training images from the learned features (Dinsdale et al., 2021). Here, instead of learning features that perform well on the worse domain, we simply maximize performance on all the training data but ensure that features do not contain any information of one particular domain.

In Sanner et al. (2021), we explore both these directions in an attempt to increase the generalization of hippocampus segmentation models, which are affected by distribution shifts such as the age of the patient and the presence of neuropsychiatric disorders, among other factors. We train with labeled data

from two domains and try to generalize to a third one. For segmentation, we employ a U-Net architecture that we augment with a VREx loss term, domain predictor, or both.

Using three openly available hippocampus segmentation datasets, we carry out experiments across scenarios with different data availability constraints: (1) in a purely supervised setting where the model works with labeled data from two domains (2) in a semi-supervised setting where the model is trained with all three datasets but only very few annotated samples from the last one; and finally (3) in a setting where only a few examples from each dataset are annotated.

In the supervised scenario, all approaches improved the performance on a third yet-unseen domain when compared to training the U-Net with both datasets without using a generalization mechanism. However, the effect we saw was either small (in cases where the base performance was already acceptable) or the gap to the intra-domain performance remained so significant that actually employing the method in a clinical workflow would be questionable. Solely utilizing the V-REx produced the most consistent improvement, also in the second semi-supervised setting.

## 6.2. Adapting data to the training domain

Let us now imagine a more restrictive scenario where we *only have labeled samples from one domain*. Here, we would need to train the supervised model solely with data from that *source domain*; and would be unable to introduce knowledge into possible domain shifts that may occur during deployment. We can even think of some situations where the model is *locked* and cannot be re-trained.

One strategy for extracting high-quality predictions from new distributions is to *transfer* these to the source domain (as visualized in Figure 6.1). For this process to be effective, we need to disentangle – either explicitly or implicitly – *domain-specific* factors from diagnostically relevant *content* such as anatomical properties of the ROI and its surroundings. We can learn a *one-to-one* mapping between the source and target domains, *one-to-many* (Mansour et al., 2008) or, ideally, *many-to-many* (Yang et al., 2019) mappings.

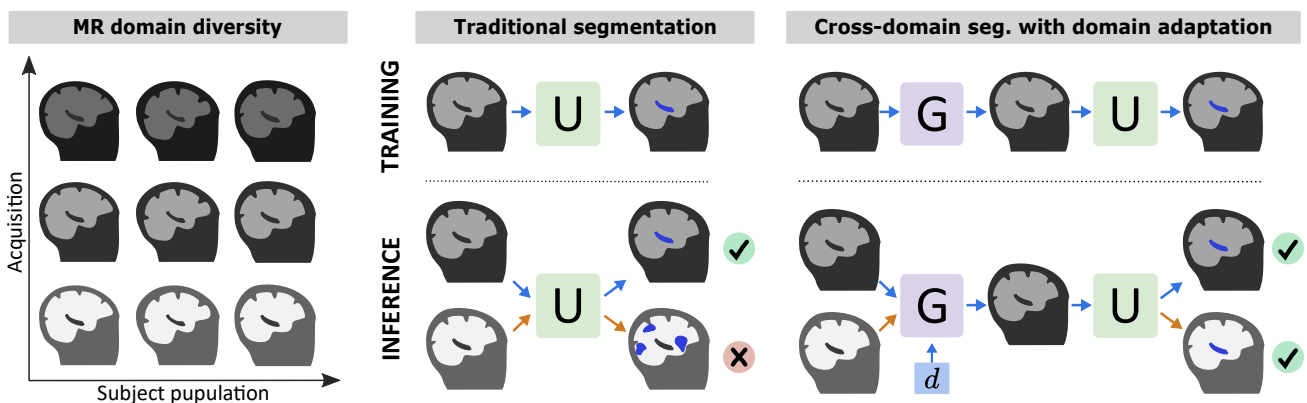


Figure 6.1.: Left: Differences in acquisition practices and patient population cause significant variability in MRIs. Right: With traditional segmentation, a model  $U$  fails to produce an acceptable hippocampus segmentation for an OOD image. If the image is transferred to the training domain with generator  $G$ , a high-quality segmentation is obtained.

---

In Kalkhof et al. (2022), we evaluate several methods that pursue such disentanglement or direct transfer between domains for the problem of hippocampus segmentation. That is to say that images are first transferred to the source domain and later passed through a downstream segmentation U-Net. We review *Dr-GAN* (Tran et al., 2017), *SD-Net* (Chartsias et al., 2019) and *SA-GAN* (Chen et al., 2018). Additionally, we propose a new method called *CDD-GAN* especially designed for image segmentation that employs a *cycle consistency loss* (Zhu et al., 2017).

While we note an improvement in the target-domain performance, a central limitation of all methods is the significant overhead they introduce in terms of computational requirements and/or training time. Additionally, again, inter-domain results are far below those of in-distribution evaluations.

We also find that a necessary step to come close to the in-distribution performance is to fine-tune the U-Net with the same labeled training data after going through the domain-transferring process. That is, converting images from the source domain to, again, the source domain. This introduces slight changes in the images that allow the extraction of higher-quality segmentations for other data. The fact that the U-Net needs to be retrained constrains the settings where the evaluated methods can be applied, basically excluding their use on pre-trained models or situations where the user no longer has access to all the training data.

When considering current regulations for clearing DL-based software, which we will review in Chapter 13, it is often the case that the predictive model is *locked* and cannot be trained post-approval. Domain adaptation may pose a reasonable alternative by transferring images to the source domain, extending the lifespan of the system without modifying the actual model. But only insofar as the model does not need to be altered post-deployment.

---

### 6.3. Conclusions and outlook

---

Learning expressive features that generalize well to unseen domains is an objective that, if fulfilled, could greatly mitigate the risks of deploying DL systems in the wild. Particularly when there is training data from several domains at once – be it labeled or not – the model can be trained to anticipate what shifts it may encounter in the future.

However, though current methods *do* improve performance on the unseen domain, a considerable gap remains to the intra-domain performance, at least for medical image segmentation tasks. This puts into doubt whether such methods would actually be useful in practice. If typical Dice segmentation performance for a certain ROI is around 90% Dice, a 70% Dice segmentation will have as little use as a 50% Dice one.

Besides, generalizability improvements often come at the cost of increased training times and computational constraints. In addition, either the architecture and/or the training objective need to be modified to accommodate this additional goal.

I thus believe that in many situations, particularly clinical settings, effectively detecting failures and limiting the use of the model to high-quality and confident predictions is the most practical avenue for managing distribution shift. This can, of course, be used alongside mechanisms to increase generalization that are deemed effective.



**Part II.**

**Continual Learning**



---

## 7. The Continual Learning Landscape

---

*Continual* or *lifelong learning* can have many meanings. Knowing what elements of the data change over time, whether these changes happen abruptly and how much knowledge we have of the data-generating process are all aspects that help us select an appropriate method. We begin this chapter by defining key concepts and establishing the notation we use in this work (Section 7.1). We then categorize continual learning scenarios and describe the setting that we believe is most relevant for medical imaging (Section 7.2). Afterward, we explain how to evaluate performance and introduce popular metrics (Section 7.3). Finally, we give an overview of strategies proposed for classification and segmentation and contextualize our work within the larger landscape of continual learning research (Section 7.4).

---

### 7.1. Key definitions

---

For the sake of simplicity, let us consider a classic supervised computer vision situation where each sample is a  $(x, y)$  tuple of an image  $x \in \mathcal{X}$  and label  $y \in \mathcal{Y}$ . In continual learning, models are trained with  $N$  tasks. Each task  $\mathcal{T}_i$  is a set of samples  $(x, y) \in \mathcal{T}_i$ , naturally divided into *train* and *test* sets.

Tasks  $\{\mathcal{T}_1 \dots \mathcal{T}_N\}$  arrive sequentially, and each is only available for a certain time interval. Our model  $\mathcal{F}_\theta : x \rightarrow \hat{y}$  is trained following this order. We denote a model trained *only* with  $\mathcal{T}_i$  as  $\mathcal{F}_i$ , and one trained with tasks up to and including  $i$  as  $\mathcal{F}_{[1, \dots, i]}$ . With  $\mathcal{F}_{\{1, \dots, N\}}$ , we refer to the upper bound of a model trained statically with the shuffled training data from all tasks.

Our goal is to train  $\mathcal{F}_{[1, \dots, N]}$  in such a fashion that it performs well on all seen tasks. The challenge here lies in preserving the knowledge acquired in the early stages, which is to say maintaining performance close to  $\mathcal{F}_{[1, \dots, i]}$  for all tasks  $\{\mathcal{T}_i\}_{i \leq N}$ , without affecting the plasticity of  $\mathcal{F}$  to learn new information. We thus wish to find a parametrization  $\theta$  of  $\mathcal{F}_\theta$  that minimizes our loss  $\mathcal{L}$  over all stages (Eq. 7.1).

$$\arg \min_{\theta} \sum_{i=1}^N \mathbb{E}_{(x, y) \sim \mathcal{T}_i} [\mathcal{L}(\mathcal{F}_\theta(x), y)] \quad (7.1)$$

This distinguishes continual from *transfer learning*, where we are only interested in maximizing performance on the last task by leveraging knowledge found in earlier stages and are not interested in knowledge accumulation/retention. The field is also closely related to *online learning*, though the focus of online learning lies in training with small batches (often even with one data point at a time) in environments with constrained resources. There is usually no focus on data drift, so the performance is measured with respect to one test set. When we speak of *lifelong learning* we refer to the practice of updating a DL-based product through its lifecycle, possibly leveraging continual learning techniques. That is to say that we consider the wider deployment setting. However, these terms are often used interchangeably in the literature (Chen and Liu, 2018).

## 7.2. Properly characterizing a continual setting

Possibly the most important aspect for defining our setting lies in identifying what changes occur in our data distribution. Based on this information, van de Ven et al. (2022) and Hsu et al. (2018) identify three continual learning scenarios.

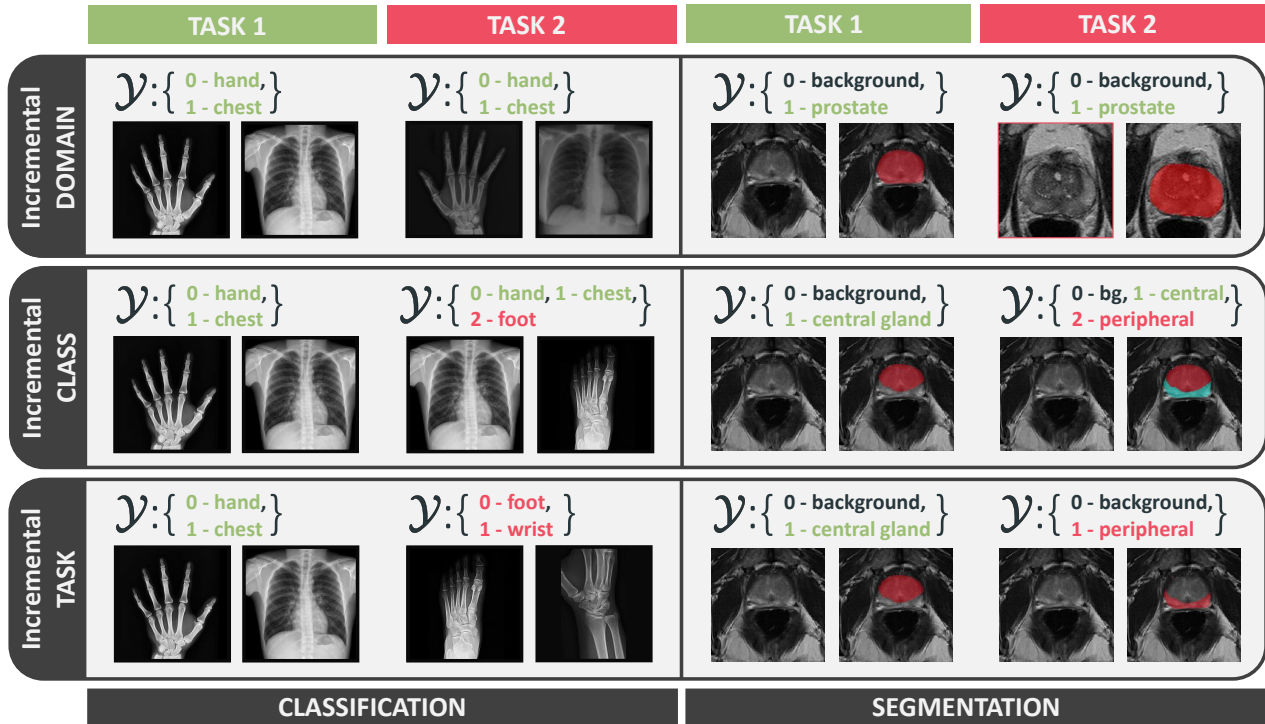


Figure 7.1.: The way tasks vary from one another determines the *continual learning scenario*. It is crucial to identify the correct scenario in order to choose an appropriate learning strategy. *Domain-incremental learning* (first row) is common in diagnostic radiology and happens when tasks are different due to discrepancies in image properties caused by the acquisition process.

Changes in image acquisition practices and in the captured content cause shifts in the image space  $\mathcal{X}$ . For instance, we previously identified variations in the scanner model or patient population as prevalent for CTs and MRIs (Chapter 1). If we only have this shift, but the problem our algorithm solves – that is to say, the label  $y$  for each  $x$  – remains the same, we are performing *domain-incremental learning*. I find this to be the most prevalent problem in medical imaging, as it affects *all deep learning models* that are deployed in dynamic environments. It is therefore also the setting I focus on in my research.

This scenario is closely related to the domain drift problem explored in previous chapters. If our models were not as susceptible to changes in the domain, we could train them sequentially without any interference. The first row in Figure 7.1 illustrates two examples in the medical domain: a classifier that must cope with contrast changes in X-rays and a model that learns to incrementally segment data from different scanners.

The second scenario, *class-incremental learning*, describes a situation where the label space  $\mathcal{Y}$  grows over time. Examples of this, visualized in Figure 7.1 (second row) include adding an additional anatomy label to our classifier or segmenter (such as the *peripheral zone* besides the *central prostate gland*). This

---

can also involve a change in the image space  $\mathcal{X}$  to accommodate the new class – as in our first example – or not – as in our second one. The main challenge here lies in dealing with a shifted label distribution, i.e. if we no longer have sufficient representation of certain classes, the optimization process will stop assigning such labels.

Finally, in *task-incremental learning* the model learns to solve fully different tasks, and the core structure of the problem changes over time. This may involve shifts in the image space, label space, or both. This scenario is closely related to *multi-task learning*, with the additional challenge that tasks arrive sequentially. The goal here is to leverage similar content for tasks that have certain semantic aspects in common. Ideally, the algorithm will learn expressive representations that allow the model to effectively leverage large amounts of data beyond that acquired for one specific problem. Real-world examples include learning to play different sports or musical instruments jointly (van de Ven et al., 2022).

### 7.2.1. Task identity and boundaries

Another way of interpreting the three continual learning scenarios, and a central aspect when categorizing a learning setting, is whether *task identity is known during training and testing*. Naturally, in task-incremental learning, the algorithm must know which task to train and perform inference on (van de Ven et al., 2022).

This is not always the case for *domain-incremental learning*. We can easily imagine a situation in which this information is present, such as a model trained with multi-institutional or multi-scanner data where we have access to the acquisition metadata. However, once we picture a more realistic clinical scenario where there are different degrees of change affecting the image distribution – such as demographic and disease patterns, acquisition conditions, and reconstruction algorithms, just to name a few – we see that these labels may not be easily obtainable or even well-defined. Additionally, data protection regulations may prevent the use of patient-identifying information in, for instance, teleradiology settings.

Whether or not we have task identity information is primarily interesting from a practical perspective: if we do have such information, we may maintain *task-specific parameters*. During training, we may set these depending on the task at hand and only update *shared parameters* in a continual manner. At inference time, we can use the task identity of the test image to build an appropriate model from our shared and task-specific weights. In the simplest case, we could maintain a different model per task. Each model would only leverage task-specific data (though it could be initialized and fine-tuned from a previous state) but suffer no forgetting. *This should always be a baseline in a scenario with task identity information.*

What parameters to share depends on the problem and architecture. A common practice is to maintain different *heads*, i.e. keep the last layers of the network task-specific as visualized in Figure 7.2, whereas earlier layers that are believed to capture more global information are shared. Other works propose keeping separate *Batch Normalization* layers, as these encode domain characteristics (Rebuffi et al., 2017a; Karani et al., 2018).

In the absence of task identity labels, another aspect to consider is whether *the boundaries between tasks are clearly defined*. That would be the case if, for instance, our model received training samples from three different scanners, one after the other, without any metadata as to what scan an image was taken with but with knowledge of *when* we start receiving data from a new scanner. In that case, we can employ an *oracle* at test time that infers the task identity for each image, as we will see in Section 8.1.1.



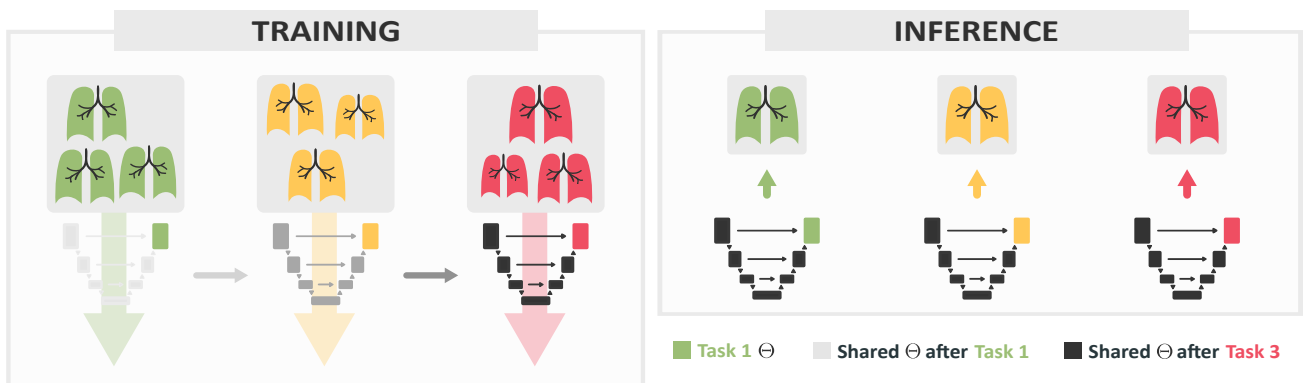


Figure 7.2.: Illustration of a classic multi-head continual learning setup. Most layers of the model are shared and updated sequentially during training. During inference, the model is constructed from the shared parameters and the task-specific head.

The situation changes when there are *slowly occurring changes in the data distribution*. This happens, say, when there are several factors that naturally shift over time related to acquisition practices and/or the patient population. In this – more complex – setting, task labels may not even be clearly defined. Yet we find this to be a particularly interesting scenario as it more closely approaches the situation in the open world, where there are several factors of change operating that we may not even be aware of.

### 7.3. Quantifying continual performance

The goal of continual learning is to train a model sequentially over an array of tasks in such a way that the final model reaches good performance *across all seen tasks*. This means that – unlike in transfer learning – the model should preserve knowledge learned early on, i.e. sustain sufficient *rigidity*. At the same time, it should be able to adapt to changes in the environment, i.e. remain *plastic*. A reasonable upper bound in terms of rigidity or knowledge preservation is the model state right after training each task, and the upper bound in terms of plasticity is sequential training with no forgetting prevention.

From a more optimistic perspective that considers the possibility of learning expressive representations that leverage all the training data, the actual upper bound in both cases is a *static or joint* training setting where all training data is merged together. If tasks really can benefit from the information contained in the other stages, this setting should display better test-time performance *over all seen tasks*<sup>1</sup>.

Figure 7.3 provides an overview of recommended metrics for continual learning. We typically measure forgetting through its inverse, *backward transfer (BWT)*, that – for any task  $\mathcal{T}_i$  except the last – subtracts the performance of model  $\mathcal{F}_{[1,\dots,i]}$  after training with  $\mathcal{T}_i$  from the performance of  $\mathcal{F}_{[1,\dots,N]}$  at the end of training with the entire sequence. Similarly, we quantify plasticity for tasks  $\{\mathcal{T}_i\}_{i>1}$  through *forward transfer*, which we can calculate by subtracting the performance of  $\mathcal{F}_i$  trained only with  $\mathcal{T}_i$  from that of model  $\mathcal{F}_{[1,\dots,i]}$  trained sequentially with all tasks up to  $\mathcal{T}_i$ . FWT can also be understood as the *performance of the model on future tasks* (Díaz-Rodríguez et al., 2018).

BWT and FWT are calculated separately for each task and averaged over the sequence. They will usually display negative values, with positive results indicating an advantage over keeping separate models per

<sup>1</sup>In practice, factors such as class unbalance and under-representation of various characteristics may prevent this.

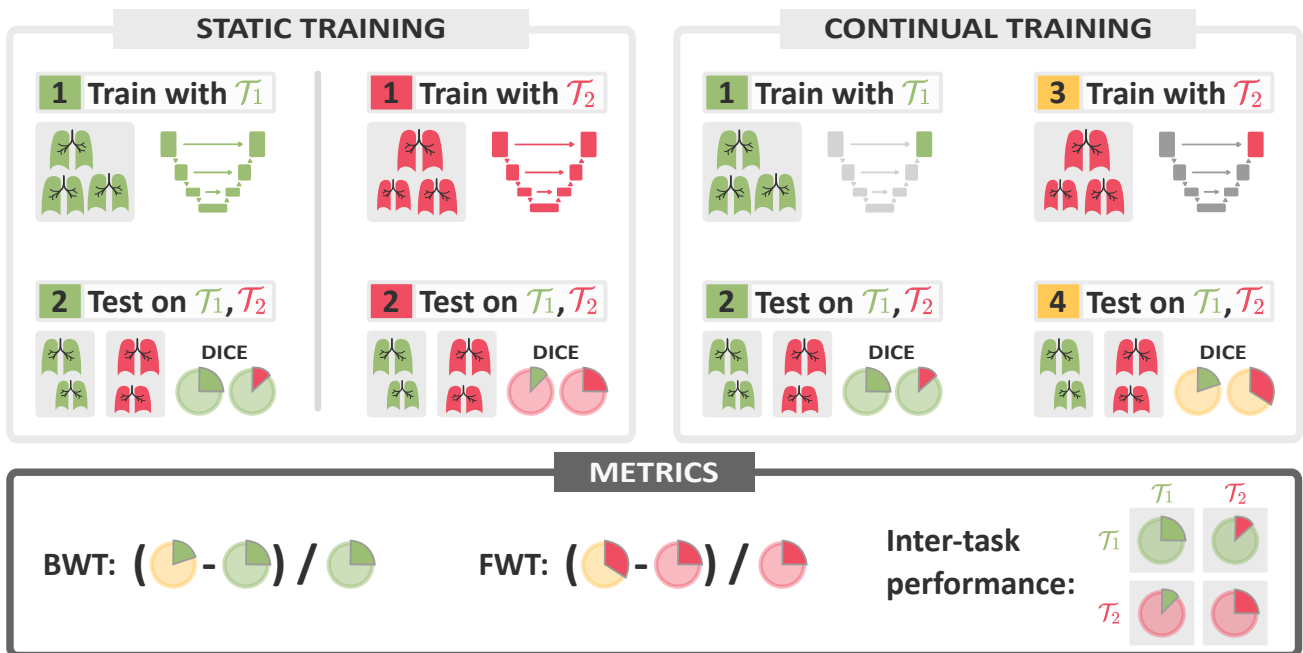


Figure 7.3.: Evaluation metrics for continual learning. Backward transfer (BWT) is, for task  $\mathcal{T}_i$ , the performance difference between the model after training with  $\mathcal{T}_i$  and after training with the following tasks. A negative BWT shows that the model *forgot* some of the information learnt in  $\mathcal{T}_i$ . Forward transfer (FWT) measures the difference between training the model with  $\mathcal{T}_i$  from scratch vs. after fine-tuning from previous tasks. A positive value shows that the model leveraged previous knowledge, as is expected for transfer learning. Negative values typically happen when a method for forgetting prevention reduces the plasticity of the model. The inter-task performance puts continual results into context by quantifying how well a static model works on other tasks.

task. We can contextualize both metrics by dividing the results by the second part of the subtraction in order to compare problems with different base performances, such as the Dice score of the left vs. right ventricular blood pool segmentation.

In addition, and particularly when we are not working with well-researched computer vision benchmarks but with more complex medical imaging problems, we should calculate the *inter-task performance* of individually trained models. This allows us to put task transferability into context. As we will see later on, two tasks  $\mathcal{T}_i$  and  $\mathcal{T}_j$  may be so similar that we see unexpected patterns such as the model performance on  $\mathcal{T}_i$  unexpectedly recuperating after training with  $\mathcal{T}_j$ . Inter-task matrices let us understand this behavior.

Though it may seem redundant, I would like to stress the importance of evaluating the performance on the test sets from all tasks individually. In addition, dataset splits should be defined once and maintained throughout the study. In an optimal case, we would repeat all experiments for all possible *tasks orderings*. However, this is often computationally unfeasible. If so, I recommend observing the inter-task performance and selecting an order that realistically mimics a slowly changing data distribution, i.e. an order that maximizes transferability between subsequent tasks.

## 7.4. Continual learning methods

In this section, we give a brief overview of popular continual learning strategies. We broadly classify approaches into the following categories, illustrated in Figure 7.4.

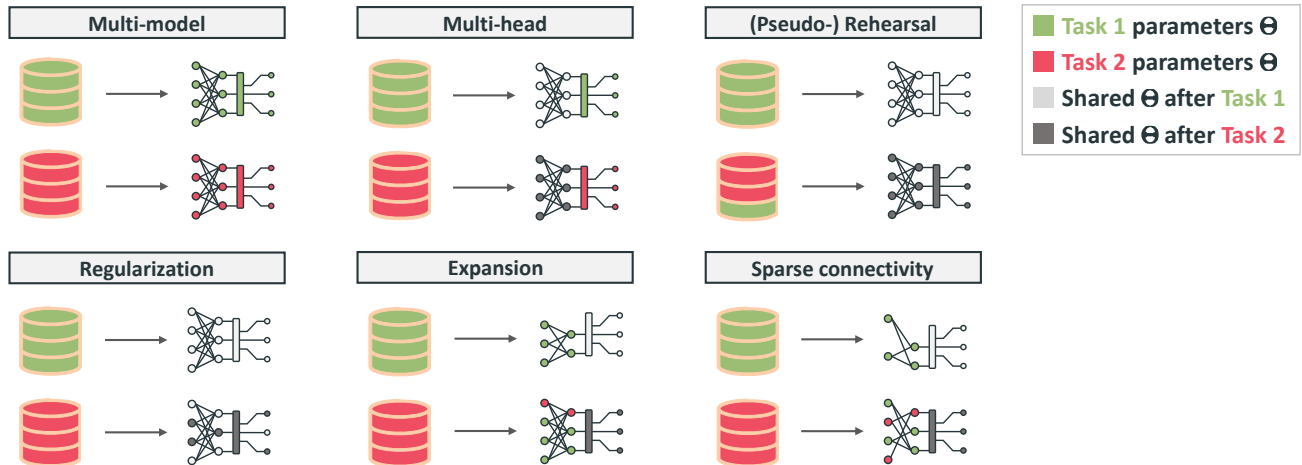


Figure 7.4.: Commonly used continual learning strategies, as well as architectural changes that are sometimes employed when training models in a continual fashion. Different approaches can be combined to build a suitable learning strategy. Some methods make a distinction between *shared* and *task-specific* parameters, displayed here with different node colorings.

**Rehearsal** methods involve storing a selection of examples from previous tasks and incorporating them into training at regular intervals (Ratcliff, 1990; Rebuffi et al., 2017b). These methods are effective in practice but can become impractical as the number of tasks increases, at which point the main technical challenge consists of selecting a representative set of samples. But the main drawback of rehearsal is that it is not admissible in any setting that restricts the storage of samples from previous stages due to – among other factors – data privacy considerations. Rehearsal-based strategies have been tested for medical images by Ozdemir et al. (2018) for humerus and scapula segmentation, Perkonigg et al. (2021) for chest CT classification, and Venkataramani et al. (2019) for x-ray lung segmentation.

**Pseudo-rehearsal** methods have been proposed as viable rehearsal alternatives precisely for situations where samples cannot be stored. This category encompasses both **generation**-based and **distillation**-based strategies. The first involves synthesizing samples *similar* to those seen in early training stages (Draeos et al., 2017; Shin et al., 2017; Ostapenko et al., 2019; Rao et al., 2019). Here, caution should be taken so that no actual inputs can be restored from the generated ones (Yu et al., 2019). The second research direction proposes distilling knowledge from a *teacher* to a *student* model, for instance by extracting outputs from the last model state with samples from the new task and trying to maintain the outputs close during fine-tuning (Li and Hoiem, 2017; Lee et al., 2019).

**Regularization** methods assess the importance of each parameter after each training stage for the corresponding data; and prevent these from being modified too heavily in later stages. This is implemented by adding an additional term to the loss function. By tuning how much this term is weighted, an acceptable trade-off can be obtained between knowledge preservation and model plasticity. Different metrics have been proposed for calculating the importance. In the medical domain, the popular *Elastic Weight Consolidation* (Kirkpatrick et al., 2017) method was evaluated for chest X-Ray lesion classification

---

(Lenga et al., 2020), as well as glioma (van Garderen et al., 2019) and white matter lesion segmentation (Baweja et al., 2018) and Özgün et al. (2020) adapt *Memory Aware Synapses* (Aljundi et al., 2018) to brain segmentation. Zhang et al. (2021) define an importance metric using domain knowledge for prostate segmentation in MR images.

**Sparse-connectivity** approaches take a more direct route and avoid interference by separating the pathways used by different tasks (Knoblauch et al., 2014; Goodrich and Arel, 2014; Ellefsen et al., 2015). Some reserve certain neurons or connections for each task and turn them on or off during inference (Mallya and Lazebnik, 2018; Golkar et al., 2019). This directly prevents forgetting, but also limits the network’s capacity for features that are not shared. Therefore, sparsity-based approaches are best reserved for cases where the model is thought to be over-parametrized. In addition, such methods invariably require knowledge of task identity.

**Expansion** strategies follow a similar idea to sparsity-based ones. However, they *add* trainable parameters and pathways as time goes on instead of restricting existing ones. This means that performance is no longer capped by model capacity. So-called *network-growing* methods change the model architecture by physically adding new parameters (Terekhov et al., 2015; Wang et al., 2017; Draelos et al., 2017; Yoon et al., 2018), whereas others keep sets of task-specific parameters that are interchanged depending on the task at hand. A common strategy consists of learning task-specific Batch Normalization layers (Rebuffi et al., 2017a; Karani et al., 2018), as tested for brain MR segmentation by Karani et al. (2018).

Before selecting a subset of methods to take for a particular problem, we recommend considering the following constraints and requirements of the continual learning scenario:

- **Is it possible to store a subset of training samples?** Are there no data privacy or storage constraints that would prohibit this? In that case, a **rehearsal**-based method would likely perform best. A sample selection strategy should be employed that uses the memory buffer efficiently.
- **Must the model architecture and loss remain unchanged?** Involving a new objective can have consequences on model performance. **Rehearsal** methods or **generative pseudo-rehearsal** approaches that artificially synthesize new data avoid this issue. **Expansion** methods can also be utilized as long as they involve maintaining *interchangeable* task-specific parameters.
- **Is there a specific trade-off desired between plasticity and rigidity?** In that case, select a strategy that adds a loss term for knowledge preservation, such as **regularization** or **distillation**.
- **Is persistent memory a limiting factor?** This could be the case in certain resource-constraint settings. In that case, rehearsal and expansion-based methods should be avoided. **Sparsity** or some forms of **regularization** could be good options.
- **Is GPU memory a limiting factor?** This is more likely, particularly when working with medical images. We have found pseudo-rehearsal approaches that either generate synthetic inputs or use a distillation term to be very computationally heavy. Certain regularization approaches can also be problematic. For settings where GPU access is limited, we recommend **sparsity** or **expansion**-based methods (for the latter, only those that exchange instead of add parameters).
- **Is the scenario task-agnostic?** Or are the boundaries between tasks and task identities given? Most methods proposed for task-agnostic learning are **rehearsal**-based (Aljundi et al., 2019a,b; Jin et al., 2021; Perkonigg et al., 2021; Srivastava et al., 2021), though other approaches can be adapted to this setting by employing a task-selection oracle (Aljundi et al., 2017). As described in Section 7.2.1, task precedence information allows us to maintain *task-specific parameters*.

---

## 8. Expansion Methods and Task-agnostic Learning

---

Expansion-based methods, particularly solutions that do not change the model architecture but instead *interchange* parameter states, are a flexible yet often overlooked continual learning avenue. In this chapter, I present the work I did in this direction and explain why I believe it is, at the moment, the best solution for adapting DNNs to dynamic clinical environments.

---

### 8.1. The papers

---

We discussed in previous sections how continual learning attempts to adapt the model to changes in the environment while preserving previous knowledge. However, there is a simple way to avoid catastrophic forgetting or a loss in model plasticity: maintaining different model states, one for each task. This is an alternative that *should always be considered in continual evaluations*. Of course, it implies that we know from which task each sample originates.

In this section, I present ways to adapt expansion-based methods to task-agnostic settings; first by introducing an *oracle* that infers task identity during inference (Section 8.1.1), and then by proposing a method that leverages OOD detection to handle slowly shifting task boundaries (Section 8.1.2).

#### 8.1.1. What is wrong with continual learning in medical image segmentation?

I began working on this manuscript at the beginning of my doctoral studies with Georgios Sakas and Anirban Mukhopadhyay, and we uploaded a version to *arXiv* on October 21<sup>st</sup>, 2020. After a review process lasting more than a year, the paper was rejected due to having only results for a prostate use case, relying on in-house data, and the segmentation architecture not being state-of-the-art. We recently carried out a significant revision, where we report results using only openly available data for three anatomies, namely prostate, hippocampus, and right ventricle. For the experiments, we make use of our *Lifelong nnU-Net* project extending the *nnU-Net* pipeline. I had the support of Nick Lemke as student assistant for running the experiments, who is now a co-author.

The work is the result of having tested many popular continual learning methods in medical image segmentation and obtaining disappointing results. The main point we try to make is that several methods employ task-specific components, implying a continual learning scenario with knowledge of task labels, yet the performance in the final model falls below maintaining one model-per task. It is inspired by the work of Aljundi et al. (2017), who proposes an *oracle* based on autoencoder networks that supply class identity information even when this is not given. We adapt that approach to image segmentation and develop an evaluation strategy that proposes appropriate multi-model baselines for each continual learning scenario.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."





Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**What is wrong with continual learning in medical image segmentation?**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "What is wrong with continual learning in medical image segmentation?" (González et al. 2023) is currently under review. It constitutes a joint work of Camila González, Nick Lemke, Georgios Sakas and Anirban Mukhopadhyay.*

*This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].*

*As corresponding and leading author, C. González led the overall research design, literature review and writing process of the paper. The choice of methodological framework and experimental setup were likewise made by C. González. The implementation of the code was performed by C. González and N. Lemke. N. Lemke conducted the experiments. C. González and N. Lemke contributed to the analysis of the data and results. The methodology, results and discussion were mainly written by C. González. G. Sakas provided his counsel on the usability of the generated segmentation tasks for several downstream tasks and contributed to the database. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor, who also contributed with continuous feedback during all phases of the paper writing process. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum:	01 / 30 / 2023	01 / 30 / 2023	01 / 30 / 2023	01 / 30 / 2023
Unterschrift:				
	Camila González	Nick Lemke	Georgios Sakas	Anirban Mukhopadhyay

# What is wrong with Continual Learning in Medical Image Segmentation?

Moving beyond catastrophic forgetting and towards practical knowledge accumulation.

Camila González<sup>1,\*</sup>, Nick Lemke<sup>1</sup>, Georgios Sakas<sup>1,2</sup>, and Anirban Mukhopadhyay<sup>1</sup>

<sup>1</sup>Technical University of Darmstadt, Karolinenpl. 5, 64289 Darmstadt, Germany

<sup>2</sup>MedCom GmbH, Dolivostraße 11, 64293 Darmstadt, Germany

\*camila.gonzalez@gris.tu-darmstadt.de

## ABSTRACT

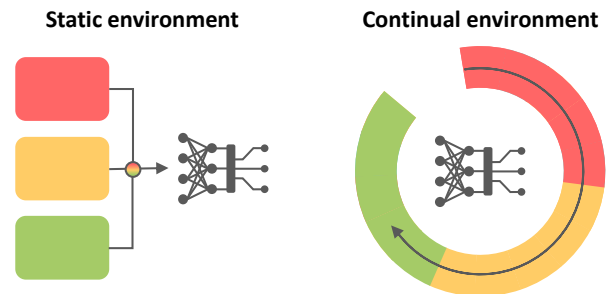
Continual learning protocols are attracting increasing attention from the medical imaging community. In continual environments, datasets acquired under different conditions arrive sequentially; and each is only available for a limited period of time. Given the inherent privacy risks associated with medical data, this setup reflects the reality of deployment for deep learning diagnostic radiology systems. Many techniques exist to learn continuously for image classification, and several have been adapted to semantic segmentation. Yet most struggle to accumulate knowledge in a meaningful manner. Instead, they focus on preventing the problem of *catastrophic forgetting*, even when this reduces model plasticity and thereon burdens the training process. This puts into question whether the additional overhead of knowledge preservation is worth it – particularly for medical image segmentation, where computation requirements are already high – or if maintaining separate models would be a better solution. We propose *UNEG*, a simple and widely applicable multi-model benchmark that maintains separate segmentation and autoencoder networks for each training stage. The autoencoder is built from the same architecture as the segmentation network, which in our case is a full-resolution nnU-Net, to bypass any additional design decisions. During inference, the reconstruction error is used to select the most appropriate segmenter for each test image. Open this concept, we develop a fair evaluation scheme for different continual learning settings that moves beyond the prevention of catastrophic forgetting. Our results across three regions of interest (prostate, hippocampus, and right ventricle) show that UNEG outperforms several continual learning methods, reinforcing the need for strong baselines in continual learning research.

## 1 Introduction

Supervised deep learning in a static setup is considered the de-facto standard for benchmarking the performance of learning-based medical image segmentation systems. In the static setup, an annotated dataset is divided into training, validation and testing subsets onto which the performance of the learning system is evaluated. Before making this division, all available data is shuffled to ensure that the samples are identically distributed. Yet this is not realistic for diagnostic radiology<sup>1</sup>. To obtain robust medical imaging models, it is necessary to leverage data from a variety of sources and continue learning over time. But two constraints may arise when handling medical data due to privacy regulations. These are that (a) data must be stored in predefined servers and so multiple datasets cannot be shuffled, and (b) some of it is only available for training the model during a limited period.

The medical imaging community is becoming aware of the discrepancies between how deep learning algorithms are evaluated and the performance drop that occurs in real clinical settings. This manifests in an increased interest in alternative training and evaluation protocols such as federated<sup>2-4</sup> and continual learning<sup>5-15</sup>. Continual learning in particular addresses the possibility that datasets from different domains arrive sequentially and are only accessible during a certain time interval, as illustrated in Figure 1.

The main technical challenge of continual learning is the prevention of *catastrophic forgetting*, which occurs when deep



**Figure 1.** Static vs. continual protocols for training and evaluating machine learning models. Each color represents a different domain. In static settings, all data is shuffled and used at once to train the model. In continual environments, the model acquires knowledge over time and can cope with losing availability to part of the data.

learning models adapt too strongly to idiosyncrasies in the last data batches and lose the ability to handle domains seen in the initial stages of training. Most existing continual learning approaches look to prevent this loss, and their evaluations focus on this aspect. Yet the actual goal of continual learning should be achieving *positive backward transfer*, which occurs when the performance on data from early domains *improves* as training continues. In a real clinical setting, only an approach that is successful in this second goal would be

deployed. Otherwise, maintaining a separate model for each data source would be preferred.

Particularly for the problem of medical image segmentation, where large model architectures are used and computational requirements are already high due to the dimensionality of the data, we find multi-model approaches to be a practical solution. Of course, we do not always have access to *domain identity labels* during inference. That is to say, we may not have any knowledge on the origin of a particular image. The question is then raised of how to select the most appropriate model. For this, we propose a simple approach based on image autoencoders. By training an autoencoder network per task, we can select the segmenter corresponding to the lowest reconstruction error.

The *Expert Gate* method has been previously proposed for image classification using autoencoders that reconstruct features extracted from an *AlexNet* network<sup>16</sup>. Considering the architectural similarities between autoencoder and segmentation networks, in that both produce an output with the same dimensionality as the input, we instead suggest *replicating the segmentation architecture for the autoencoder*, and merely adapting the last layer and training objective to the mean squared error. This poses the advantages that (1) the architecture is already optimized for the problem at hand, (3) the process of data preparation and pre-processing, which has likewise been tuned for the specific problem, can remain the same and (3) no additional design decisions are required. We refer to this multi-model solution as the *U-Net Expert Gate*, or **UNEG**. We show that this is a better strategy for the segmentation problem with our results across three different regions of interest (ROIs), namely prostate, hippocampus, and right ventricle, using the state-of-the-art *nnU-Net* pipeline<sup>17</sup>.

Despite several strategies being introduced to permit continual learning, no article has, as of yet, properly introduced the variants and related terminology of continual learning in the medical imaging segmentation context and proposed an **evaluation scheme that moves beyond forgetting prevention** for each setting. This makes it difficult to compare different approaches and assess their potential usability in clinical practice. A major goal of this article is to provide a holistic view regarding the trade-offs and terminologies of existing strategies, as well as the differences between continual learning scenarios. We propose fair multi-model benchmarks to compare against new continual learning approaches that take these into consideration. In this way, we aim to establish a common ground to facilitate discussion around continual learning in the coming years.

Our contributions are as follows:

1. Introducing a fair multi-model benchmark for continual learning in medical image segmentation; a solution that can be easily used alongside highly specialized models. The method uses an *oracle* to select the appropriate model in situations where (a) domain identity information is not provided or (b) the system should handle observations from previously-unseen sources.
2. Proposing an autoencoder-based oracle that follows the same architecture as the segmentation network, which is already suitable for the current image modality and region of interest, thereby requiring no additional design decisions.
3. Showing the effect of catastrophic forgetting in three magnetic resonance imaging (MRI) segmentation tasks, namely prostate, hippocampus, and right ventricle, and how this can be avoided.

We start this work by formalizing the problem of continual learning and describing existing scenarios in section 2. We give an overview of related work on continual learning and its applicability to image segmentation and medical data in section 3. In section 4, we describe our proposed evaluation setup based on a multi-model approach and a domain identification strategy with image autoencoders. We report results for three different MR segmentation problems in section 6. Finally, we give an outlook for future research in section 7.

## 2 Problem formulation

We start this section by introducing key terminology and a taxonomy for continual learning settings based on a) how data distributions from different sources differ and b) whether domain identity information is available during inference. We then motivate the need for specialized continual learning solutions to prevent catastrophic forgetting. Finally, we propose a new way to evaluate continual learning approaches that moves beyond forgetting prevention.

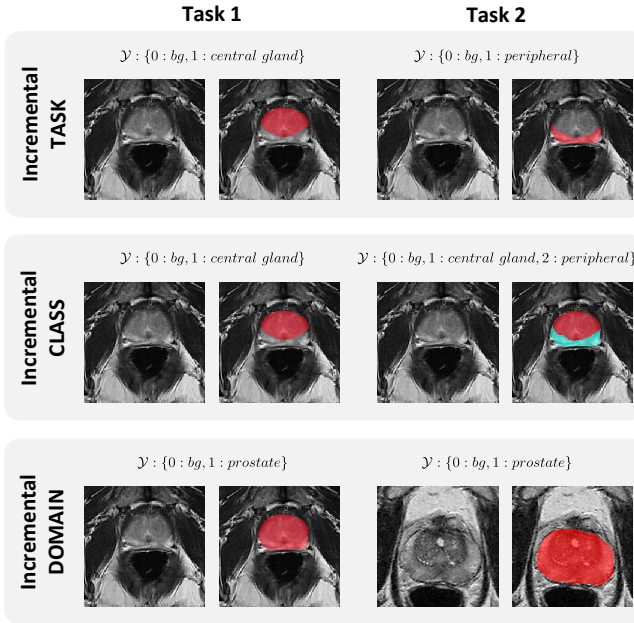
### 2.1 Continual Learning

In a continual learning setting, the database consists of  $N$  datasets  $\{X_i, Y_i\}_{i \leq N}$ . Each dataset comprises samples  $(x, y)$  where  $x$  is an input and  $y$  is the corresponding annotation. In this work we focus on image segmentation, so each sample is a pair of an image  $x$  and segmentation mask  $y$ . The goal is to train a model  $\mathcal{F}$  that performs well on all datasets, which arrive sequentially and are only available for a limited period. Figure 1 illustrates how data is received in a continual learning protocol. Model  $\mathcal{F}$  is trained sequentially with  $\{X_1, Y_1\}$ ,  $\{X_2, Y_2\}$  and so on until it has acquired information contained in all available data sources. Unlike in a static setup, samples  $(x, y) \in \{X_1, Y_1\}$  share certain characteristics resulting from the generation process which are not present in samples from other training stages.

Differences in the sample distributions can manifest in various ways. van de Ven et al.<sup>18</sup> introduce three continual learning scenarios. These are demonstrated in Figure 2 for the case of prostate segmentation on T2-weighted MRI.

In the *incremental task scenario*, both the input and label spaces can vary, and usually task identity information is provided. For the exemplary task, the network would segment the prostate central gland for task 1, and the peripheral zone for task 2. This scenario is not seen frequently for image





**Figure 2.** Three continual learning scenarios, exemplified for the case of prostate segmentation. In the incremental task scenario, both the image and label spaces can widely differ, but task identity information is available during inference. The incremental class problem consists of sequentially expanding the number of classes that a model can handle. Finally, in the incremental domain setting, the label space remains the same but there are differences in the image space, often resulting from the acquisition protocol, and we have no access to domain identity labels.

segmentation. In the *incremental class scenario*, new classes are incrementally added to the label space. In our example, whilst only the class for the central gland can be learned with dataset 1, an additional class for the peripheral zone is learned with dataset 2, and the final model can segment both classes. This scenario is interesting to explore, but only meaningful under certain specific circumstances, and adapting the architecture is required each time a new class is introduced. Finally, in the *incremental domain scenario*, the knowledge learned is the same semantically, but there are differences in terms of image characteristics, that is to say that  $X_i \simeq X_j$  and  $Y_i = Y_j$ . We argue that this is the most prevalent scenario in practice for medical imaging, as it comes into play each time a system must learn incrementally from images acquired from different sources and/or at different times. We thus focus on this scenario for the rest of this work.

A *domain*  $\mathcal{D}$  is a set of image characteristics that are particular to the acquisition source but independent of the content of interest. For instance, images obtained with one MR machine may have a different contrast than those obtained with another, and the ROI may be captured from a slightly different angle depending on the acquisition protocol.

We further differentiate between whether domain identity information is available during inference, as this is ambiguous in the incremental domain scenario. Based on this, we define three settings. In the simpler case, that we name *Domain Knowledge*, test inputs have the form  $(x, i)$ , where  $i$  specifies that  $x \in X_i$ . This would be the case if, for instance, a model is trained with data from three different scanners, and we receive at inference time metadata on the scanner used to acquire each image. In the second scenario, *No Domain Knowledge*, no such information is available during testing. The main advantage of having domain identity information is that *domain-specific parameters* can be maintained that are not shared across domains, so only a subset of the model parameters, the *shared parameters*, must be trained in a continual fashion. For classification, the feature extraction part of the model is typically shared whereas the last network layers are domain-specific and set during inference depending on the domain precedence of the test instance.

There are two limitations on the usability of the *Domain Knowledge* scenario. Firstly, the model can only be applied to data from domains that have been observed during training. Secondly, it is not realistic to assume that domain labels will be available during deployment. There are a lot of variabilities in how information regarding image acquisition is encoded in the metadata of image files, even within the same institution. In certain cases, such as in teleradiology systems, this information may not be available due to the anonymization process. Additionally, in a realistic dynamic setting where multiple factors vary over time, it is not trivial to assess which of those factors is the most relevant.

## 2.2 The Catastrophic Forgetting debacle

The naïve way to train a model continuously is to perform sequential fine-tuning, executing training steps as data arrives. But if samples belonging to distinct datasets stem from different distributions, this violates the assumption that data be i.i.d. as required by stochastic gradient descent. One prevalent degree of variability for diagnostic radiology are particularities in the image domains that arise from the acquisition protocols or equipment vendors used for each dataset, a phenomenon commonly known as *domain interference* or *domain shift*<sup>19</sup>.

Neural networks trained using stochastic gradient descent on sequentially arriving data adapt too strongly to domain properties present in the last batches. For data similar to that seen in the initial stages of training, this causes a significant drop in performance known as *catastrophic forgetting*<sup>20</sup>. If, instead, the model is protected so that it does not change too much, it is possible that future knowledge cannot be acquired. Therefore, special attention must be taken during training to ensure that the final model performs well on data similar to that seen at *any stage of training*. However, we argue that simply preventing forgetting and ensuring model plasticity are only the first objectives of continual learning.

### 2.3 Existing approaches to mitigate Catastrophic Forgetting and their evaluation

Different strategies have been developed to reduce the degree of forgetting. The applicability of several popular continual learning methods to medical imaging has also been explored in the past, both for classification<sup>8,12,15,21</sup> and segmentation<sup>6,7,9–11,13</sup>. These methods are compared against the baseline of performing sequential fine-tuning and other continual learning approaches, as well as against the upper bound of performing static training that would be preferred if all data were available at once.

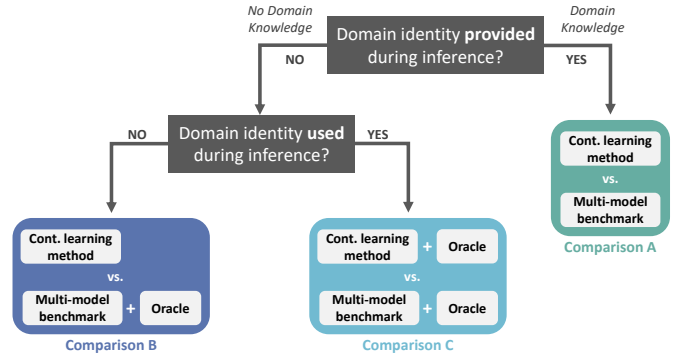
However, we find that often these methods are not compared against the simple multi-model benchmark, i.e. maintaining one model per domain, which could potentially outperform the proposed approach. They *are* compared to the static setup of training a model with all, *joint*, data. Yet unlike a multi-model solution, this is a clear upper bound that is not applicable in a continual learning setting.

In the *Domain Knowledge* case, where domain information can be used at test time, a naïve way to avoid catastrophic forgetting is to maintain a separate model  $\mathcal{F}_i$  for each domain  $\{X_i\}_{i \leq N}$ . Model  $\mathcal{F}_1$  is only trained with data  $\{X_1, Y_1\}$ ,  $\mathcal{F}_2$  is initialized with the parameters of  $\mathcal{F}_1$  and fine-tuned with  $\{X_2, Y_2\}$ , and so on. Of course, only model  $\mathcal{F}_N$  leverages all available data, as it is sequentially fine-tuned with all data batches. However, no catastrophic forgetting takes place. Each model  $\mathcal{F}_i$  maintains the same performance it had after training on data  $\{X_i, Y_i\}$ . Additionally, models do not suffer from a loss of plasticity as the training process does not discourage large parameter changes.

We found only three methods<sup>11,13,22</sup> in medical imaging to report a similar comparison. In all cases, either the multi-model solution is treated as an upper bound or the improvement against it is minimal. In addition, some publications report a *backward transfer* or *forgetting* measure that compares the performance at the end of training with all domains to that at the end of training with the corresponding one<sup>10,12,14,23</sup>. The results only rarely show a situation of *positive backward transfer* or *negative forgetting*, where the performance improves after learning from other domains.

This is worrying, as it questions the applicability of the proposed methods in real clinical workflows. If a model does not improve after continuing training, what advantage comes in maintaining a single model, other than reduced storage requirements? Particularly in clinical practice, space constraints are rarely an issue, and robustness is always the priority.

Despite this, it is common practice to assume domain identity information during testing. In the *No Domain Knowledge* scenario or if the model should be applicable to previously-unseen domains, **many continual learning methods cannot be used directly.**



**Figure 3.** Flowchart for the proposed evaluation of continual learning methods. If domain identity information is available, the method should be compared against a benchmark containing one model per domain (comparison A). If domain information is not available, or the model should be usable for data of previously-unseen domains, then it should instead be compared against a combination of the multi-model benchmark and an oracle that identifies which model to apply. If the method requires such information, then the same oracle can be used in conjunction (comparison C). Otherwise, the method is evaluated without using domain knowledge (comparison B). If the benchmark outperforms the newly proposed method, the latter is only an improvement under additional problem-specific constraints, such as limited memory.

### 2.4 Proposed evaluation / Our solution

We argue that all methods proposed for continual learning should be compared against a *multi-model benchmark*, as it is a straightforward solution to the catastrophic forgetting problem and thus the trivial lower bound for performance. While other continual learning methods require a particular way of training the model or architectural adjustments, the benchmark does not alter the training procedure or architecture of the main model.

In Figure 3, we illustrate how an ideal evaluation would proceed. **If domain identity information can be used during inference**, a comparison can take place with the benchmark. If domain information is unavailable, we propose using an *oracle* to infer the closest domain for an incoming image  $x$ . **If the proposed method requires domain information**, then a comparison can take place for both the continual learning method and the benchmark using domain information inferred by the oracle. **If the method does not use domain information**, then only the benchmark would make use of the oracle.

If the proposed method outperforms the benchmark, the evaluation can go forward. If it performs worse, then its use is only reasonable under additional constraints, such as limitations on persistent memory.

### 3 Related work

In this section, we give an overview of common continual learning strategies and exemplary methods. We also summarize recent research on continual learning for image segmentation and diagnostic radiology.

#### 3.1 Strategies to prevent catastrophic forgetting

Methods to prevent catastrophic forgetting can be broadly classified into the following strategies.

**Rehearsal** methods store a subset of examples from previous tasks and periodically interleave these during training<sup>24–26</sup>. They perform best in practice but do not scale well to an increasing number of domains and can only be used if there are no restrictions against storing training data. **Pseudo-rehearsal** approaches simulate the same effect without requiring the storage of data files. Within this category are methods that use long-term and short-term memory components<sup>27,28</sup>, generate examples similar to those of previous tasks with generative models<sup>29–32</sup> or use distillation losses to encourage the outputs of the latest model to remain close to previous outputs<sup>33,34</sup>. Despite no data being explicitly stored, care should be taken when using generative methods that the models not be sufficient to regain the data<sup>35</sup>.

**Sparse-connectivity** methods discourage overlap between representations learned while training with different tasks<sup>36–38</sup>, based on the theory that representational overlap causes catastrophic forgetting. Some methods directly reserve a certain portion of the network for each task. Yet unused network regions are masked, and inference takes place with all regions up to that of the current task<sup>39,40</sup>. Therefore, the performance does not decrease. A disadvantage of this strategy is that the capacity of the network is limited for features that are not shared, and task identity is always required.

**Network growing** strategies prevent the loss of model capacity by continuously adding new trainable parameters. Some maintain a single network, to which additional layers or neurons are added as new tasks appear<sup>41,42</sup>. This is especially successful when complemented by rehearsal training<sup>29,43</sup>. Others train a separate model per task and combine these by merging the parameter values<sup>44,45</sup> or learning connections between the models<sup>46</sup>. Alternatively, one model state is chosen during inference<sup>16</sup>. The main disadvantages of network-growing approaches are that the space requirements grow linearly with the number of tasks, and network architectures must be continuously adapted.

**Regularization** approaches calculate an importance value for each parameter after training a model with data for domain  $\mathcal{D}_i$ , and penalize the divergence from those parameters, weighted by the importance, when training with data of domains  $\mathcal{D}_j : j > i$ . Methods differ mainly on how they assess the importance<sup>47–49</sup>.

**Bayesian** methods have also been developed to reduce catastrophic forgetting in Bayesian Neural Networks<sup>50–52</sup>. The disadvantage is that training networks in a Bayesian manner comprises a considerable time overhead.

Other strategies include learning **domain-invariant features**<sup>53</sup> or maintaining **different batch normalization parameters**<sup>54</sup>.

#### 3.2 Applications to semantic segmentation

Most work on continual learning has focused on the classification problem. However, recent research has looked into adapting these strategies for semantic segmentation.

Following the pseudo-rehearsal strategy, a simple approach consists of saving image statistics of previous domains for pseudo-example generation<sup>55</sup>. Using a distillation loss, Shmelkov et al.<sup>56</sup> prevent catastrophic forgetting for object detection. Michieli et al.<sup>57</sup> expand on this by distinguishing between output and feature-level distillation terms, through results show that considering the divergence of intermediate features rarely improves the performance. Recently, Cermelli et al.<sup>58</sup> introduce a distillation loss that takes into account how the proportion of background pixels change across domains and show that this improves segmentation performance.

Beyond pseudo-rehearsal methods, Nguyen et al.<sup>23</sup> propose a regularization-based approach that uses saliency maps as a measure for parameter importance, and Matsumoto and Yanai<sup>9</sup> introduce a sparse-connectivity method that learns task-specific masks.

Unlike in classification or regression, where even wrong outputs may seem plausible, semantic segmentation poses the additional challenge that output masks must maintain certain characteristics to resemble the ground truth, such as having a certain number of connected components or adhering to geometric properties. Sequential learning causes the integrity of masks to deteriorate, even if the correct ROI is identified. This is reflected in a greater gap between the performance of static and continual learning results in semantic segmentation.

#### 3.3 Continual learning in medical imaging

Several works have explored the applicability of continual learning methods to medical imaging, mostly adapting existing regularization and pseudo-rehearsal strategies to the task at hand. In medical image segmentation, the research is mostly focused on brain MRIs.

Lenga et al.<sup>8</sup> evaluate both *EWC* and *LwF* for the problem of Chest X-Ray lesion classification. *EWC* has also been evaluated for glioma<sup>7</sup> and white matter lesion<sup>5</sup> segmentation. As is often the case with using this approach, the level of catastrophic forgetting is decreased but learning new domains becomes more difficult. Özgün et al.<sup>10</sup> also adapt the *MAS* regularization method to brain segmentation. The authors slightly modify how the importance is calculated and normalized, which causes a small improvement over the regular *MAS* implementation.

Ozdemir et al.<sup>11</sup> look at how best to select previous examples to prevent catastrophic forgetting using a rehearsal method for the task of segmenting humerus and scapula bones on MR images. A rehearsal method with dynamic memory is also evaluated for the problem of chest CT classification<sup>15</sup>. Venkataramani et al.<sup>13</sup> explore continual lung segmentation

on X-Ray images using a memory component that stores data samples for each target domain.

Karani et al.<sup>22</sup> mitigate the performance loss in brain MR image segmentation obtained with different scanners or scanning protocols. The method consists of learning a U-Net with shared convolutional layers but domain-specific batch normalization layers; and performs slightly better than training a separate network for each domain. However, it requires data from several domains to be available at once.

Some works focus on learning domain-independent features or learning transformations between feature spaces. Kim et al.<sup>21</sup> aim to directly create a domain-independent feature space by maximizing the mutual information between the feature space  $Z$  and output space  $Y$ . This is achieved through minimizing the  $L2$  distance between features  $z$  and the reconstruction  $h(g(z))$ , where  $G : X \implies Y$  and  $h = h^{-1}$ . The proposed method outperforms *EWC* and *LwF* in the classification of tuberculosis from chest X-Rays, as well as on *CIFAR10* and *CIFAR100*. Elshahawy et al.<sup>14</sup> use an adversarial approach to disentangle domain-dependent from domain-independent features. Their method outperforms *LwF* in the incremental class learning setting. Ravishankar et al.<sup>12</sup> instead propose using feature transformer networks that turn features extracted for each domain appropriate for using with a following classification network. They show positive results for X-Ray pneumothorax and ultrasound cardiac view classification. However, no comparison takes place between the transformed features and those trained for each task.

Finally, the Bayesian *Distributed Weight Consolidation* method is proposed for performing brain segmentation in a distributed manner for an ensemble of networks trained with Variational Inference<sup>6</sup>.

Due to the complexity of the semantic segmentation problem and the geometric differences between ground truth masks from different datasets, methods are mostly evaluated with very similar datasets and only one ROI. In this work, we report results for the segmentation of the prostate, hippocampus, and right ventricle.

## 4 Methods

If we store the model state after each training stage, there are three ways to select model parameters depending on whether domain knowledge is provided and/or used. The flowchart in Figure 3 depicts how this translates to different evaluation settings, and Figure 4 provides a graphical portrayal of how a model state can be selected.

In the *Domain Knowledge* setting where identity labels are given (upper image), the model state right after training with the corresponding domain can be used. If this information is not given, as in the second image, an *oracle* can be used at test time to select the most appropriate model. Finally, single-model methods use the state after finishing training with all domains for extracting all predictions.

Every continual learning method can be compared to a *multi-model benchmark*. As the name suggests, the bench-

mark maintains one model per domain. It, therefore, poses a fair comparison where no catastrophic forgetting or loss of plasticity take place. In the *Domain Knowledge* scenario, simply using the appropriate model state according to the domain of a test sample is a fair comparison. In *No Domain Knowledge*, the previous method is an upper bound, and an *oracle* must be used to select the best state at test time.

The main drawbacks of the benchmark are that the space requirement grows linearly with the number of domains and that the model  $\mathcal{F}_i$  does not leverage information from  $\{X_j, Y_j\}_{j \geq i}$ . There is therefore no possibility of positive backward transfer, though also no forgetting in the individual models.

For the sake of simplicity, we assume that *all* the model parameters are taken from the selected state. In practice, it is possible that only certain layers (such as the end of the decoder) are kept domain-dependent, and the rest are shared. For instance, many continual learning methods maintain separate model *heads*, one per domain.

### 4.1 Autoencoder-based oracle

Inspired by Aljundi et al.<sup>16</sup>, we use autoencoder networks to build our oracle for domain identification during inference.

An *autoencoder* is a neural architecture designed to reconstruct the input, i.e. learning the mapping  $\mathcal{A} : X \rightarrow X$  by minimizing the mean-squared error between the original image and the reconstruction, for  $n$  samples (Eq. 1).

$$\mathcal{L}_{MSE} = \frac{1}{n} \sum_{i=1}^n (x_i, \mathcal{A}(x_i))^2 \quad (1)$$

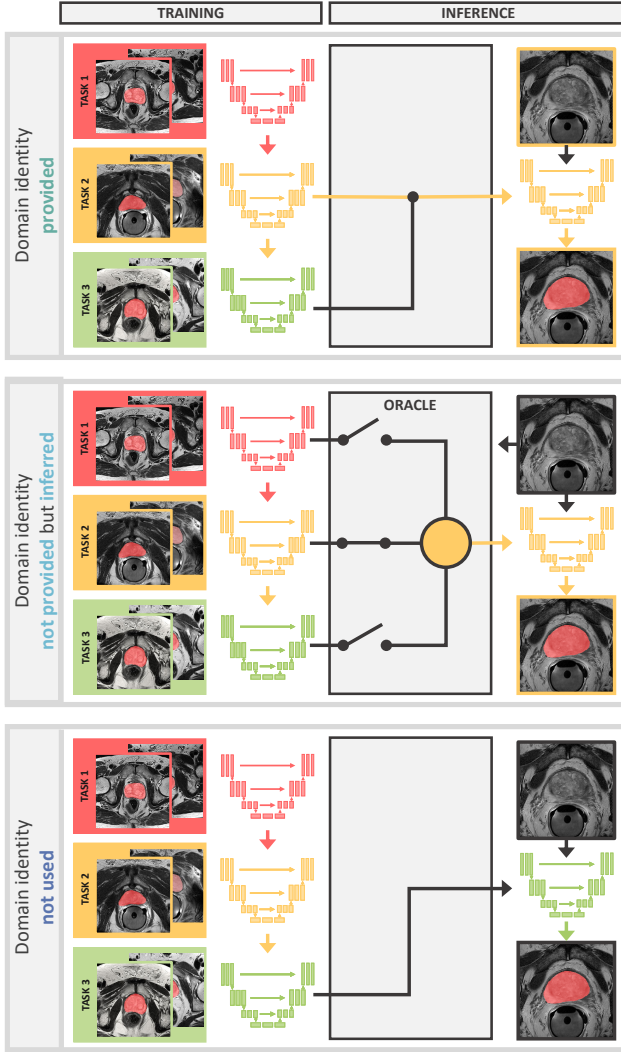
For our oracle, one autoencoder  $\mathcal{A}_i$  is trained for each incoming dataset. If the hardware allows it, this can occur in parallel to training the segmentation model for the same training stage.

When a test image  $x$  arrives, a reconstruction is then created using each of the trained autoencoders  $\{\mathcal{A}_i\}_{i \leq N}$ . The domain of the autoencoder with the smallest reconstruction error is used to segment  $x$  (Eq. 2).

$$\hat{y} = \mathcal{F}_i(x); \operatorname{argmin}_i \mathcal{L}_{MSE}(x, \mathcal{A}_i(x)) \quad (2)$$

Typically, autoencoders contain an *encoder* that reduces the spatial dimensionality of the input and a *decoder* that returns it to the initial dimensions. The method proposed by Aljundi et al.<sup>16</sup> uses a small CNN to reconstruct *AlexNet* features.

We take a different approach, leveraging the fact that autoencoders follow the same structure as U-Net architectures. We, therefore, propose replicating *the same architecture that is used for the segmentation problem*, as it is already suitable for the specific problem and does not require making any additional design decisions. The only variation we do is that we remove the last layer, which discretizes output values into prediction labels so that we can train the model with the *MSE* between the inputs and outputs.



**Figure 4.** Training and testing of continual learning methods under three different settings. During training, shared parameters are trained by sequential fine-tuning, and the state of domain-dependent parameters is saved after each training stage. In the upper image, domain information is provided. During inference, the model state corresponding to the domain of the test image is restored. If this information is not provided, it is inferred through image properties by a domain detection oracle, as shown in the middle diagram. Finally, the lower image shows the case where the continual learning method does not use any domain information.

## 5 Experimental Setup

In the following, we describe our data corpus for three different image segmentation problems. We also state details on the architecture and training procedure of the segmentation and autoencoder networks and the continual learning methods that we compare.

### 5.1 Data

We evaluate the proposed approach across three anatomies in MRIs, namely the prostate, hippocampus, and right ventricle.

For prostate, we use T2-weighted MRIs from five different sites<sup>59</sup>. The datasets are different in terms of manufacturer and acquisition settings, and each contains 12 to 30 cases. We train in the following order: *BIDMC* → *I2CVB* → *HK* → *UCL* → *RUNMC*. The delineations encompass both the central gland and the peripheral zone.

The hippocampus corpus consists of the *Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation* (henceforth referred to as *Dryad*) dataset<sup>60</sup>, the *Harmonized Hippocampal Protocol* dataset<sup>61</sup> (*HarP* for short) and the data released as part of the *Medical Segmentation Decathlon (DecathHip)*<sup>62</sup>. We train in the order *DecathHip* → *Dryad* → *HarP*. The segmentation masks cover the posterior and anterior hippocampus.

For right ventricle segmentation, we use the data released for the *Multi-Centre, Multi-Vendor and Multi-Disease (M&M) Cardiac Segmentation Challenge*<sup>63</sup>, which contains two datasets with 75 samples each, the first acquired with *Siemens* scanners, the second with *Philips*.

### 5.2 Segmentation nnU-Net

We use the patch-based, three-resolution variation of the *nnU-Net*<sup>17</sup>. One model is trained per anatomy, and we perform 250 epochs per dataset.

The architecture and training configuration, such as the patch size, are automatically configured by the framework. As we perform continual training, the settings selected for the first dataset are maintained for subsequent data of the same anatomy. The patch sizes used are [28, 256, 256] for the prostate examinations, [40, 56, 40] for the hippocampus, and [14, 256, 224] for the right ventricle.

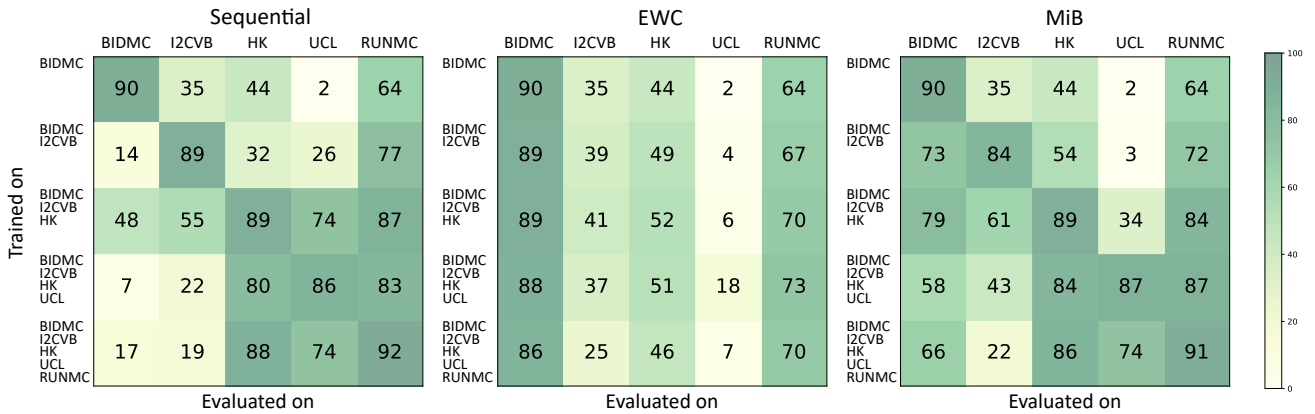
### 5.3 Continual Learning baselines

We compare our proposed benchmark to three popular continual learning methods. We use the implementations and default hyperparameters of the *Lifelong nnU-Net* framework<sup>64</sup>. All methods are trained in 3D full-resolution with the same configurations as stated in the previous section.

We explore *EWC* (Elastic Weight Consolidation)<sup>48</sup> with a  $\lambda_{EWC} = 0.4$ , *LwF* (Learning without Forgetting)<sup>33</sup> with a temperature of 2 and *MiB* (Modeling the background)<sup>58</sup>, with the alpha parameter of 0.9 and KD loss weighting to 1. These are the parameters suggested by default. We additionally compare to sequential learning without using any mechanism for knowledge preservation and to the upper bound of static, *joint* training where we use all training data at once.

### 5.4 Autoencoder architectures

The *Expert Gate* method by Aljundi et al.<sup>16</sup> proposes extracting features from a pre-trained *AlexNet model*<sup>65</sup> and reconstructing them with a two-layer CNN. We test this approach, though we do not believe it is the most suitable for medical images, and refer to this method as **AlexNet  $z$ -CNN**. Since



**Figure 5.** Mean Dice of five prostate segmentation tasks, after training with each of five stages. The diagonal from the upper left to lower right corners shows the score after training with the corresponding training data.

AlexNet is trained for RGB image classification, we use two-dimensional slices. Replicating the channel 3 times and feeding it through the AlexNet results in a feature volume of 256 channels with reduced spatial resolution.

As autoencoders follow a similar architecture to U-Nets and other popular segmentation models, we propose mimicking the same architecture as used for segmentation. In the case of the nnU-Net, each model is already configured for the specific particularities of the data. We simply modify the last layer to not discretize the logits with a softmax function; and minimize the reconstruction error to the input instead of the segmentation loss. This is our proposed **UNEG** (*U-Net Expert Gate*) oracle. We also try an alternative autoencoder that directly reconstructs the input images, namely that offered by the **MONAI** framework<sup>66</sup> which consists of convolution, instance normalization and PReLU blocks. Different to the nnU-Net autoencoder, there is no change in spatial resolution.

To assess whether it is the features or the model which are the most relevant, we experiment as well with using *features from the corresponding nnU-Nets*, reconstructed with a 2-layer CNN autoencoder. We call this method, which also works in three dimensions, **nnU-Net z-CNN**. Similarly, we try out a **CNN** network to reconstruct the images directly. In both cases, the features are taken from the last decoder block.

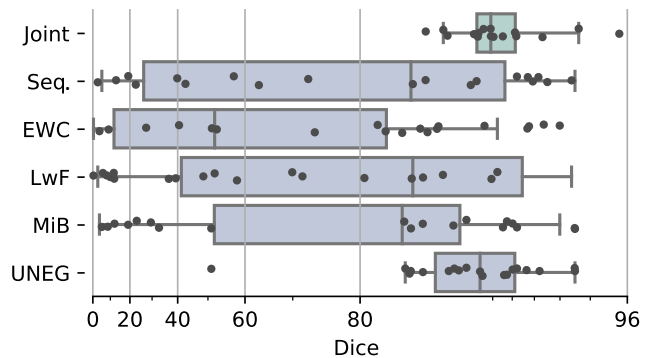
All autoencoders are trained to minimize the MSE between the reconstruction and the input for 250 epochs. During inference/evaluation, we use the segmentation network that corresponds to the autoencoder with the smallest reconstruction error.

## 6 Results

We first compare the proposed **UNEG** benchmark to several continual learning methods. We then perform an ablation study where we explore alternatives for the autoencoder oracle. Finally, we look at a few visual examples of reconstructions from the nnU-Net autoencoders.

### 6.1 Comparison to continual learning methods

Figure 6 visualizes the results for prostate segmentation. The first boxplot shows the upper bound of a model trained statically with all data. For prostate segmentation, a Dice of around 90% is expected in a static scenario. We then see the results for training a model sequentially without forgetting prevention, where the scores are distributed across the performance spectrum. Three continual learning methods follow, namely EWC, LwF and MiB, the latter of which performs best. The multi-model **UNEG** benchmark performs considerably better than continual learning approaches, though there is still a gap in performance to static training.



**Figure 6.** Dice scores of the final model state on test data from five prostate segmentation datasets.

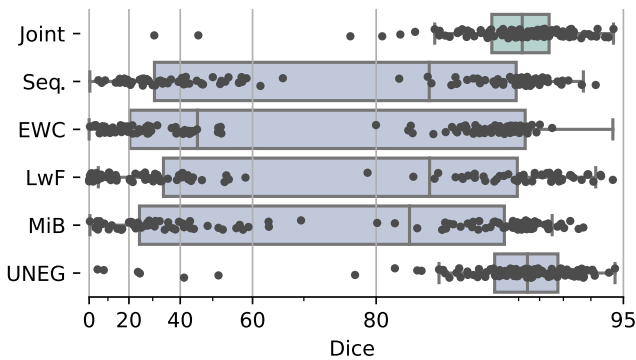
Figure 5 provides additional insight into how the methods perform at different stages. For regular sequential training, we notice the performance deterioration typical of catastrophic forgetting. *MiB* reduces this effect somewhat, but the final model still produces low-quality segmentations for the first few stages. *EWC* instead displays a different behavior: the performance remains high for the first task, but the model is clearly constrained in its ability to capture new knowledge.

A similar but more pronounced behavior takes place for hippocampus segmentation. Figure 7 visualizes these results. For

**Table 1.** Ablation study on the selection of the best autoencoder architecture for an oracle that infers task identity. We report the mean Dice, BWT<sup>64</sup> and the accuracy at selecting the “correct” task identity.

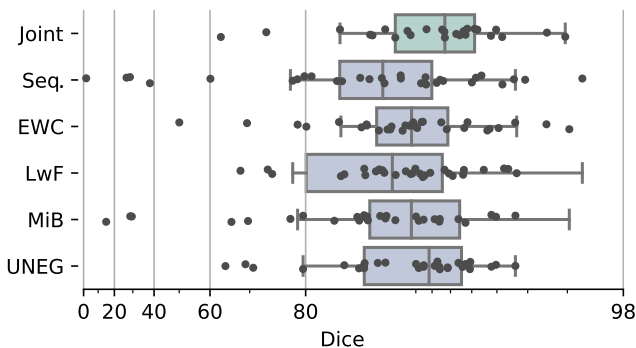
Method	Prostate			Hippocampus			Right ventricle		
	Dice $\uparrow$	BWT (%) $\uparrow$	Acc. $\uparrow$	Dice $\uparrow$	BWT (%) $\uparrow$	Acc. $\uparrow$	Dice $\uparrow$	BWT (%) $\uparrow$	Acc. $\uparrow$
Task identity	89.6 $\pm$ 2.0	-	-	90.6 $\pm$ 2.2	-	-	89.6 $\pm$ 1.5	-	-
AlexNet $z$ -CNN	87.2 $\pm$ 6.6	-3.5 $\pm$ 6.0	84.2	73.3 $\pm$ 19.2	-23.9 $\pm$ 23.9	24.1	81.1 $\pm$ 9.9	-19.2 $\pm$ 0.0	50.0
UNEG	87.0 $\pm$ 6.5	-3.8 $\pm$ 5.9	68.4	89.0 $\pm$ 3.1	-2.6 $\pm$ 2.6	97.4	88.7 $\pm$ 2.3	-2.0 $\pm$ 0.0	90.0
MONAI	66.2 $\pm$ 29.7	-30.0 $\pm$ 36.2	31.6	65.0 $\pm$ 22.8	-29.4 $\pm$ 29.4	13.8	89.6 $\pm$ 1.5	-0.0 $\pm$ 0.0	100.0
CNN	80.3 $\pm$ 11.3	-11.5 $\pm$ 13.0	52.6	72.0 $\pm$ 28.3	-0.0 $\pm$ 0.0	56.0	87.5 $\pm$ 4.3	-5.6 $\pm$ 0.0	60.0
nnU-Net $z$ -CNN	84.2 $\pm$ 7.0	-3.5 $\pm$ 6.1	52.6	60.7 $\pm$ 26.4	-15.0 $\pm$ 14.1	46.6	81.1 $\pm$ 9.9	-19.2 $\pm$ 0.0	50.0

all single-model methods, we see a clear separation between samples that are correctly segmented and those for which performance is dismal.



**Figure 7.** Dice scores of the final model state on test data from the three hippocampus segmentation datasets.

For our third region of interest, the right ventricular blood pool (Figure 8), all methods perform much better. Only sequentially training the model and *LwF* display a visible performance loss when compared to the joint training upper bound. This is likely due to the fact that there are only two tasks, and the domain differences caused by using different scanners may not be as significant as those introduced in the hippocampus datasets, where the patient populations differ.



**Figure 8.** Dice for the task of right ventricle segmentation on test data from both datasets, Philips and Siemens.

## 6.2 Ablation study

We perform an ablation study where we test several autoencoder options in Table 1. We calculate the mean Dice over all tasks of the final model states, *BWT* as defined by González et al.<sup>64</sup> and how accurately the oracle properly identifies the domain.

The first row shows the upper bound of using the ground truth task identities, as is possible in the *Domain Knowledge* scenario. The second row is the CNN autoencoder proposed by Aljundi et al.<sup>16</sup>, which reconstructs *AlexNet* features. We then report the results of using the nnU-Net autoencoder (*UNEG*) and several other settings, which are described in Section 5.4. While *AlexNet z-CNN* correctly identifies the domain for most prostate cases, it fails to do so for the hippocampus and cardiac examinations. UNEG instead achieves high accuracy for the three anatomies, but most importantly a high Dice across all domains.

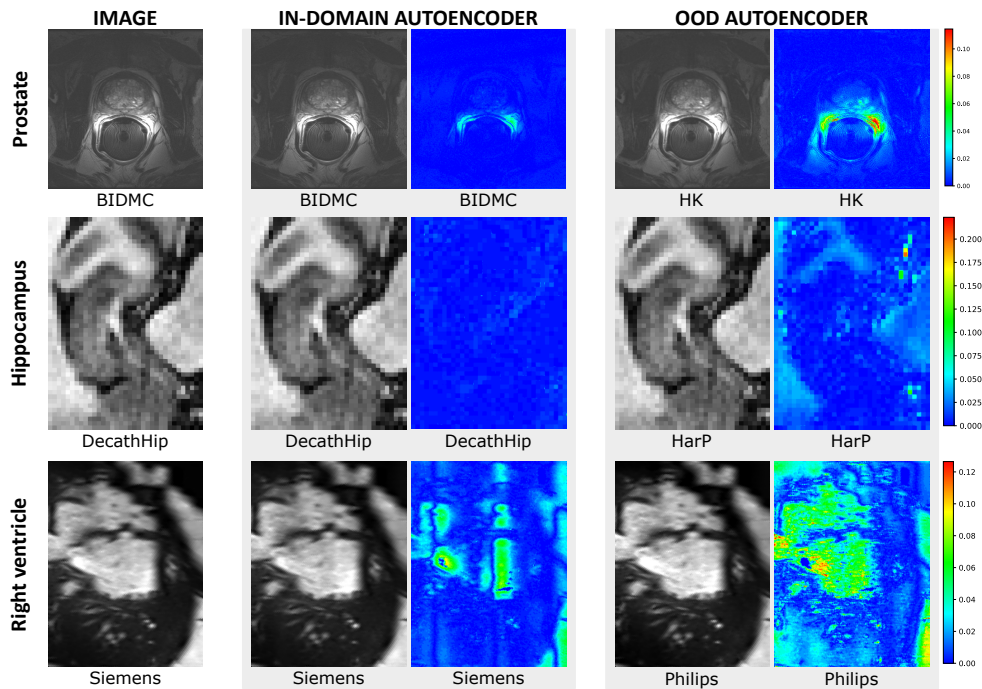
## 6.3 Qualitative evaluation of image reconstructions

We can observe exemplary image reconstructions for the three anatomies in Figure 9. The first column shows the original image. Then, we see the reconstruction produced by the autoencoder trained with data from the same domain and the residual image highlighting the differences between the two. This is followed by the reconstruction made by an autoencoder from a different domain. The reconstructed images look good at first sight, but when looking at the residual distance, we notice that the quality is much worse than for in-distribution reconstructions.

## 6.4 Discussion

The conversation on continual learning often revolves around *catastrophic forgetting*, the brusque fall in performance for domains seen in early training stages. Often, a trade-off is pursued between *rigidity* and *plasticity*, whereby the model preserves previous knowledge but can still adapt to changes in the environment. Yet there is a trivial way to circumvent the loss of performance in both directions: keeping a separate model per domain – stored after completing the respective training stage – and using the appropriate model at test time.

In this work, we introduce an evaluation workflow alongside a multi-model benchmark, a fair baseline that maintains one model per domain. The benchmark has no possibility of gaining useful information from data seen in later stages



**Figure 9.** Qualitative evaluation of the reconstructions extracted by the “correct” autoencoder vs. an autoencoder trained with a different domain. Besides the reconstructed image we display the residual between the image and the reconstruction.

for an earlier domain, i.e. no *positive backward transfer* can occur. Yet related works show us that this is rarely the case in practice, which puts into question whether the additional complexity of continual learning approaches that modify the training procedure is worth it.

Central to our evaluation scheme is the question of whether domain identity information is present during inference. Many continual learning methods assume so, which makes it possible to maintain a *multi-head* architecture where only some parameters are shared. Yet this assumption does not hold in many real-world settings, which compromises the applicability of the proposed methodologies.

For cases where identity labels are not known, we propose *UNEG (U-Net Expert Gate)*, which trains one autoencoder per domain. The autoencoder replicates the architecture of the segmentation model – in our case, a patch-based nnU-Net. This makes use of the fact that the architecture and pre-processing steps are already tuned to the particular input data and task. Our empirical evaluation exploring various autoencoder settings confirms that this is the most effective way to select the correct model at test time.

One limitation of our approach is that one additional model, namely the autoencoder, needs to be trained per stage. This can occur in parallel to training the segmenter but still implies the use of additional computation resources. In future work, we will explore more efficient *oracle* strategies.

## 7 Conclusion and Outlook

Many methods exist to prevent catastrophic forgetting for image classification, and several have been adapted with relative success to semantic segmentation. Yet **few methods achieve positive backward transfer**, i.e. while the model does not forget how to deal with data seen in early training stages, it also does not leverage information seen later on. In such cases, a multi-model solution would be preferable in clinical practice, where reliability is paramount and there is rarely a lack of persistent storage. In this work, we present a multi-model strategy alongside a fair evaluation framework for continual learning methods.

The proposed evaluation considers the fact that continual learning approaches often rely on receiving domain identity information during inference. This may not be the case in real-world dynamic environments, where metadata may be concealed for privacy reasons or the model must handle data from previously-unseen sources.

Continual learning methodologies are gaining a lot of attention from the diagnostic radiology community. Just like we are seeing more works that evaluate models on out-of-distribution data, we hope that training and evaluating models in a continual fashion and quantifying their backward transferability becomes common practice.

## Acknowledgements

This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMV11-2520DAT03A].



## References

1. Pianykh, O. S. *et al.* Continuous learning ai in radiology: implementation principles and early applications. *Radiology* 200038 (2020).
2. Brisimi, T. S. *et al.* Federated learning of predictive models from federated electronic health records. *Int. journal medical informatics* **112**, 59–67 (2018).
3. Silva, S. *et al.* Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, 270–274 (IEEE, 2019).
4. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37**, 50–60 (2020).
5. Baweja, C., Glocker, B. & Kamnitsas, K. Towards continual learning in medical imaging. *arXiv preprint arXiv:1811.02496* (2018).
6. McClure, P. *et al.* Distributed weight consolidation: A brain segmentation case study. In *Advances in Neural Information Processing Systems*, 4093–4103 (2018).
7. van Garderen, K., van der Voort, S., Incekar, F., Smits, M. & Klein, S. Towards continuous learning for glioma segmentation with elastic weight consolidation. *MIDL* (2019).
8. Lenga, M., Schulz, H. & Saalbach, A. Continual learning for domain adaptation in chest x-ray classification. In *Medical Imaging with Deep Learning* (2020).
9. Matsumoto, A. & Yanai, K. Continual learning of image translation networks using task-dependent weight selection masks. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, 129–142 (Springer, 2020).
10. Özgün, S., Rickmann, A.-M., Roy, A. G. & Wachinger, C. Importance driven continual learning for segmentation across domains. In *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11*, 423–433 (Springer, 2020).
11. Ozdemir, F., Fuernstahl, P. & Goksel, O. Learn the new, keep the old: Extending pretrained models with new anatomy and images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11*, 361–369 (Springer, 2018).
12. Ravishankar, H., Venkataramani, R., Anamandra, S., Sudhakar, P. & Annangi, P. Feature transformers: privacy preserving lifelong learners for medical imaging. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, 347–355 (Springer, 2019).
13. Venkataramani, R., Ravishankar, H. & Anamandra, S. Towards continuous domain adaptation for medical imaging. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 443–446 (IEEE, 2019).
14. Elskhawy, A. *et al.* Continual class incremental learning for ct thoracic segmentation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 106–116 (Springer, 2020).
15. Hofmanninger, J. *et al.* Dynamic memory to alleviate catastrophic forgetting in continuous learning settings. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, 359–368 (Springer, 2020).
16. Aljundi, R., Chakravarty, P. & Tuytelaars, T. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3366–3375 (2017).
17. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. methods* **18**, 203–211 (2021).
18. van de Ven, G. M., Tuytelaars, T. & Tolias, A. S. Three types of incremental learning. *Nat. Mach. Intell.* **4**, 1185–1197 (2022).
19. Glocker, B., Robinson, R., Castro, D. C., Dou, Q. & Konukoglu, E. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597* (2019).
20. McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, vol. 24, 109–165 (Elsevier, 1989).
21. Kim, H.-E., Kim, S. & Lee, J. Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I*, 520–528 (Springer, 2018).
22. Karani, N., Chaitanya, K., Baumgartner, C. & Konukoglu, E. A lifelong learning approach to brain mr segmentation across scanners and protocols. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 476–484 (Springer, 2018).
23. Nguyen, G. *et al.* Dissecting catastrophic forgetting in continual learning by deep visualization. *arXiv preprint arXiv:2001.01578* (2020).

24. Ratcliff, R. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychol. review* **97**, 285 (1990).
25. Rebuffi, S.-A., Kolesnikov, A., Sperl, G. & Lampert, C. H. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010 (2017).
26. Aljundi, R., Lin, M., Goujaud, B. & Bengio, Y. Gradient based sample selection for online continual learning. *Adv. neural information processing systems* **32** (2019).
27. Ans, B. & Rousset, S. Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connect. science* **12**, 1–19 (2000).
28. Kemker, R. & Kanan, C. Fearnert: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563* (2017).
29. Draelos, T. J. *et al.* Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 526–533 (IEEE, 2017).
30. Shin, H., Lee, J. K., Kim, J. & Kim, J. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, 2990–2999 (2017).
31. Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P. & Nabi, M. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11321–11329 (2019).
32. Rao, D. *et al.* Continual unsupervised representation learning. *Adv. Neural Inf. Process. Syst.* **32** (2019).
33. Li, Z. & Hoiem, D. Learning without forgetting. *IEEE transactions on pattern analysis machine intelligence* **40**, 2935–2947 (2017).
34. Lee, K., Lee, K., Shin, J. & Lee, H. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 312–321 (2019).
35. Yu, N., Davis, L. S. & Fritz, M. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).
36. Knoblauch, A., Körner, E., Körner, U. & Sommer, F. T. Structural synaptic plasticity has high memory capacity and can explain graded amnesia, catastrophic forgetting, and the spacing effect. *PLoS one* **9**, e96485 (2014).
37. Goodrich, B. & Arel, I. Unsupervised neuron selection for mitigating catastrophic forgetting in neural networks. In *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 997–1000 (IEEE, 2014).
38. Ellefsen, K. O., Mouret, J.-B. & Clune, J. Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS computational biology* **11**, e1004128 (2015).
39. Golkar, S., Kagan, M. & Cho, K. Continual learning via neural pruning. *arXiv preprint arXiv:1903.04476* (2019).
40. Mallya, A. & Lazebnik, S. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7765–7773 (2018).
41. Terekhov, A. V., Montone, G. & O’Regan, J. K. Knowledge transfer in deep block-modular neural networks. In *Conference on Biomimetic and Biohybrid Systems*, 268–279 (Springer, 2015).
42. Wang, Y.-X., Ramanan, D. & Hebert, M. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2471–2480 (2017).
43. Yoon, J., Lee, J., Yang, E. & Hwang, S. J. Lifelong learning with dynamically expandable network. In *International Conference on Learning Representations* (International Conference on Learning Representations, 2018).
44. Jafari, O. H., Groth, O., Kirillov, A., Yang, M. Y. & Rother, C. Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 4620–4627 (IEEE, 2017).
45. Misra, I., Shrivastava, A., Gupta, A. & Hebert, M. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3994–4003 (2016).
46. Rusu, A. A. *et al.* Progressive neural networks. *arXiv preprint arXiv:1606.04671* (2016).
47. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M. & Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 139–154 (2018).
48. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc. national academy sciences* **114**, 3521–3526 (2017).
49. Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3987–3995 (JMLR. org, 2017).
50. Ebrahimi, S., Elhoseiny, M., Darrell, T. & Rohrbach, M. Uncertainty-guided continual learning with bayesian neural networks. *arXiv preprint arXiv:1906.02425* (2019).
51. Nguyen, C. V., Li, Y., Bui, T. D. & Turner, R. E. Variational continual learning. *arXiv preprint arXiv:1710.10628* (2017).

52. Swaroop, S., Nguyen, C. V., Bui, T. D. & Turner, R. E. Improving and understanding variational continual learning. *arXiv preprint arXiv:1905.02099* (2019).
53. Ebrahimi, S., Meier, F., Calandra, R., Darrell, T. & Rohrbach, M. Adversarial continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 386–402 (Springer, 2020).
54. Rebuffi, S.-A., Bilen, H. & Vedaldi, A. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, 506–516 (2017).
55. Wu, Z., Wang, X., Gonzalez, J. E., Goldstein, T. & Davis, L. S. Ace: Adapting to changing environments for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2121–2130 (2019).
56. Shmelkov, K., Schmid, C. & Alahari, K. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE International Conference on Computer Vision*, 3400–3409 (2017).
57. Michieli, U. & Zanuttigh, P. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0 (2019).
58. Cermelli, F., Mancini, M., Bulo, S. R., Ricci, E. & Caputo, B. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9233–9242 (2020).
59. Liu, Q., Dou, Q., Yu, L. & Heng, P. A. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Med. Imaging* (2020).
60. Kulaga-Yoskovitz, J. *et al.* Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset. *Sci. data* **2**, 1–9 (2015).
61. Wisse, L. E. *et al.* A harmonized segmentation protocol for hippocampal and parahippocampal subregions: Why do we need one and what are the key goals? *Hippocampus* **27**, 3–11 (2017).
62. Antonelli, M. *et al.* The medical segmentation decathlon. *Nat. communications* **13**, 4128 (2022).
63. Campello, V. M. *et al.* Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Med. Imaging* **40**, 3543–3554, DOI: [10.1109/TMI.2021.3090082](https://doi.org/10.1109/TMI.2021.3090082) (2021).
64. Gonzalez, C., Ranem, A., dos Santos, D. P., Othman, A. & Mukhopadhyay, A. Lifelong nnunet: a framework for standardized medical continual learning. - (2022).
65. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L. & Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Inc., 2012).
66. Cardoso, M. J. *et al.* Monai: An open-source framework for deep learning in healthcare, DOI: [10.48550/ARXIV.2211.02701](https://doi.org/10.48550/ARXIV.2211.02701) (2022).

---

## Contribution and impact

A main concern that we address in the paper is the fact that very few continual learning works actually report *positive backward transfer*. This means that, while they may prevent catastrophic forgetting to some extent, the performance of the model after training with the *first* task remains best at solving it. Similarly, the sequentially fine-tuned model is better at solving the *last* task than one trained with forgetting prevention (though this is to be expected, as that model state has leveraged knowledge from preceding tasks).

Considering these objectives, if *task identity labels* were available, methods should be compared against the baseline of maintaining multi-task models. When this information is not available, then an *oracle* can be employed to infer it, and the combination of the oracle + multi-model solution is an appropriate baseline.

We base the design of our oracle on the work by Aljundi et al. (2017), who propose training one autoencoder per task. At test time, the task is used for which the respective autoencoder obtains the lowest reconstruction error. The authors reconstruct features extracted from an *AlexNet* network pretrained on *ImageNet*. Our main contribution lies in adapting this method to the medical imaging setting. We suggest that, as autoencoders and U-Net-style segmentation architectures both follow an encoder-decoder structure, the most suitable autoencoder will follow the same network architecture and pre-processing steps as the segmenter. For this, we simply remove the *softmax* layer that discretizes the outputs and replace the segmentation objective with the mean squared error. Our empirical results confirm that this is an effective strategy for the three anatomies we explore.

We carry out our evaluation in our open-source *Lifelong nnU-Net* project for state-of-the-art continual segmentation. Performing an evaluation with our new proposed benchmark thus requires minimal additional work for any researcher familiar with the nnU-Net.

## Discussion and limitations

A critique I have received several times when proposing multi-model solutions is that this is not considered *continual learning* – which seeks training *one model* sequentially – and that training several models and architectures comprises additional computational and storage concerns. I firmly disagree with the first statement. I believe continual learning describes *the constraints and characteristics of the problem scenario* and is not limited to specific aspects of the method. Regarding the second point, it disregards the fact that many pseudo-rehearsal or regularization solutions require the storage of previous model states and/or increase training overhead by growing the architectures through several heads. I also wish to highlight the difference between a multi-model approach and a deep ensemble, as in the first only one model at a time is trained, and one used for inference. This means that no more resources are needed in using a multi-model strategy other than the persistent memory required for model storage and those related to the oracle.

Nevertheless, it is true that the autoencoder-based oracle comprises an additional overhead, as a second model must be trained per task and as many forward passes are required from autoencoders as the number of tasks. In addition, we do not consider the situation where there are no clear boundaries between tasks, an additional challenge of many task-agnostic settings. We address both these considerations in the paper described in the following section.

---

### 8.1.2. Task-Agnostic continual hippocampus segmentation for smooth population shifts

The paper *Task-Agnostic continual hippocampus segmentation for smooth population shifts* (González et al., 2022d) combines my research in OOD detection with my interest in expansion-based continual learning. I presented the paper at the *MICCAI Domain Adaptation and Representation Transfer (DART)* workshop on September 23<sup>rd</sup>, 2022, where it won the best-paper runner-up award. A pre-print version is available in *arXiv* since August 5<sup>th</sup>, 2022. The publication is joint work with Anirban Mukhopadhyay, Amin Ranem, who helped adapt the evaluation code in the *Lifelong nnU-Net* code base, and Ahmed Othman, who supervised the experimental design from a neuroradiological perspective.

#### Contribution and impact

We previously stated the benefits of multi-model solutions, both as a baseline and a simple strategy to allow the continual adaptation of any DNN. One point we did not address in the past is how to deal with another characteristic of task-agnostic scenarios: the *absence of clear domain boundaries*. In this work, we look at how to deal with a slowly shifting data distribution in two different hippocampus segmentation scenarios. This adds three additional challenges to our problem: *detecting* domain shifts, keeping the *model pool small*, and selecting an appropriate model state *during inference*.

We propose *ODEx*, which uses the Mahalanobis-based method introduced in Section 3.1.2, but adapted to a setting where we do not have access to all the training data at once for estimating the distribution of network features. Key to our approach is that we select for each training stage whether an existing model is updated or a new one is initialized. Specifically, we proceed in the following fashion. We extract features from the training data (in this case, from batch normalization layers) and estimate a multi-variate Gaussian distribution. As new data comes in, we calculate the Mahalanobis distance to the distribution of all maintained model states. If the lowest distance exceeds a threshold, we initialize a new model from the closest state. Otherwise, we update the nearest model. During inference, we again calculate the distance of the image to all models and extract a prediction with the closest one.

One main contribution of our paper is that we accumulate the mean and covariance over feature stages to capture the distribution of the model as it changes over time. We find that *ODEx* reliably maintains high performance over all seen tasks, only requiring negligible additional amounts of persistent memory.

#### Discussion and limitations

One limitation of our paper is the focus on hippocampus segmentation, as the method would be well-suited to the problems where the OOD detection approach is successful. Within medical segmentation, the hippocampus has several practical advantages such as the relatively low resolution and the fact that non-expert-readers can – to a certain extent – assess the validity of the predictions.

Another aspect we do not investigate is whether only certain layers instead of the entire model could be kept task-specific. The reason we decided against this is that it introduces an additional design choice that would be dependent on the problem scenario. We believe that for maintaining the amount of persistent storage within an acceptable range as time goes on, the central aspect is *keeping the model pool small*, not storing fewer parameters per model state, which is a constant factor. Still, keeping shared parameters could potentially have advantages such as enabling positive backward transfer.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."





Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Task-agnostic continual hippocampus segmentation for smooth population shifts**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Task-agnostic continual hippocampus segmentation for smooth population shifts" (González et al. 2022) was published as a full research paper at the "Domain Adaptation and Representation Transfer (DART)". It constitutes a joint work of Camila González, Amin Ranem, Ahmed Othman and Anirban Mukhopadhyay.*

*This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].*

*As corresponding and leading author, C. González led the overall research design, literature review and writing process of the paper. The choice of methodological framework and experimental setup were likewise done by C. González. The implementation of the code were performed by C. González and A. Ranem. The central implications of this work were derived by A. Mukhopadhyay as general advisor of this work, who also contributed with continuous feedback during all phases of the paper writing process; and by A. Othman from a clinical neuroradiology perspective. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum:	01 / 09 / 2023	01 / 10 / 2023	01 / 10 / 2023	01 / 14 / 2023
Unterschrift:				
	Camila González	Amin Ranem	Ahmed Othman	Anirban Mukhopadhyay

# Task-agnostic Continual Hippocampus Segmentation for Smooth Population Shifts\*

Camila González<sup>1</sup>[0000-0002-4510-7309](✉), Amin Ranem<sup>1</sup>, Ahmed Othman<sup>2</sup>,  
and Anirban Mukhopadhyay<sup>1</sup>

<sup>1</sup> Darmstadt University of Technology, Karolinenplatz 5, 64289 Darmstadt, Germany  
[camila.gonzalez@gris.tu-darmstadt.de](mailto:camila.gonzalez@gris.tu-darmstadt.de)

<sup>2</sup> University Medical Center Mainz, Langenbeckstraße 1, 55131 Mainz, Germany

**Abstract.** Most continual learning methods are validated in settings where task boundaries are clearly defined and task identity information is available during training and testing. We explore how such methods perform in a task-agnostic setting that more closely resembles dynamic clinical environments with gradual population shifts. We propose ODEx, a holistic solution that combines out-of-distribution detection with continual learning techniques. Validation on two scenarios of hippocampus segmentation shows that our proposed method reliably maintains performance on earlier tasks without losing plasticity.

**Keywords:** Continual learning · Lifelong learning · Distribution shift.

## 1 Introduction

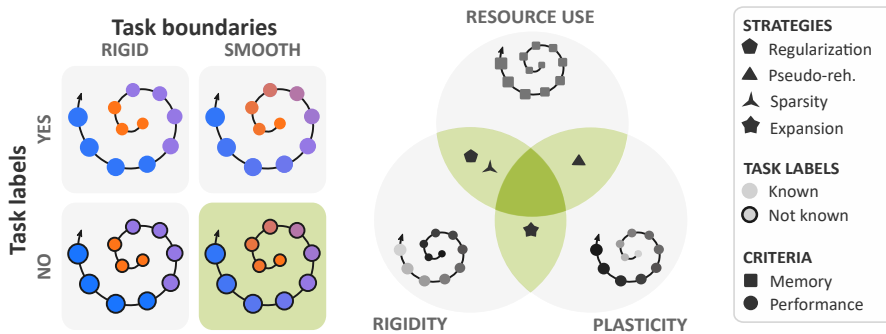
Deep learning methods are mostly validated in stationary environments where the train and test data have been carefully homogenized to preserve the i.i.d. assumption. This does not reflect the reality of clinical deployment, where acquisition conditions and disease patterns evolve over time. *Continual learning* (CL) paradigms are being explored by medical imaging researchers [19,22,27] and regulatory bodies [29] as evaluation settings that are better suited for AI in health-care. Continual methods deal with temporal restrictions on data availability by sequentially accumulating knowledge over a stream of *tasks*, each containing data from a different distribution, without revisiting previous stages.

Yet most CL approaches are validated in settings with *rigid task boundaries* and *known task labels*, which is far from how real dynamic environments behave [7]. When deviating from this simplistic problem formulation, they perform worse than simple baselines [23]. Previous research has established desirable properties for CL methods, illustrated in Fig. 1. These include no reliance on either (1) assumptions on task boundaries during training or (2) access to task identity labels, i.e. the method should be *task-agnostic* [10]. In addition, the model should (3) preserve previous knowledge while (4) maintaining sufficient plasticity to

---

\* Supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].

learn new tasks and (5) not require additional computational resources during training [7,10]. The last three objectives are often deemed to be orthogonal, i.e. most approaches either *catastrophically forget* previous knowledge (too plastic), cannot learn new tasks (too rigid) or the training time and resource requirements grow linearly with the number of tasks.



**Fig. 1.** Desiderata for continual learning [7,10]. Left: methods should not rely on rigid boundaries or task labels. Right: trade-off between plasticity, rigidity and resource use.

Methods for task-agnostic continual learning are overwhelmingly *rehearsal-based* [1,2,12,21,27], i.e. store a subset of past images or features in a memory buffer, which is not admissible in many diagnostic settings due to patient privacy considerations. *Active learning* methods also exist which rely on expert interaction [22].

Other approaches train generative models to identify distribution shifts [24] or only update the shortest sub-path of the network that allows a correct classification [6], but such solutions are computationally expensive and are therefore only evaluated in low-resolution classification settings. The field of continual learning for medical segmentation is still under-studied. Most research follows regularization-based strategies that calculate the importance of parameters and penalize their deviation [19,30]. Approaches have also been proposed for active learning [31], others allow the storage of previous samples [21,28]. Some methods leverage feature disentanglement to alleviate forgetting [16,18] or maintain task-dependent batch normalization layers [13]. To our knowledge, no method has been previously introduced for semantic segmentation that is task-agnostic and does not make use of a rehearsal component.

We propose **ODEx**, an expansion-based approach that (1) does not revisit previous stages, (2) is well-suited to a wide array of use cases, including semantic segmentation and (3) is task-agnostic, i.e. requires neither task boundaries nor task labels during training or inference. *ODEx* uses continual out-of-distribution (OOD) detection to signal when to *expand* the model and select the best parameters during inference. Although we maintain multiple parameter states in



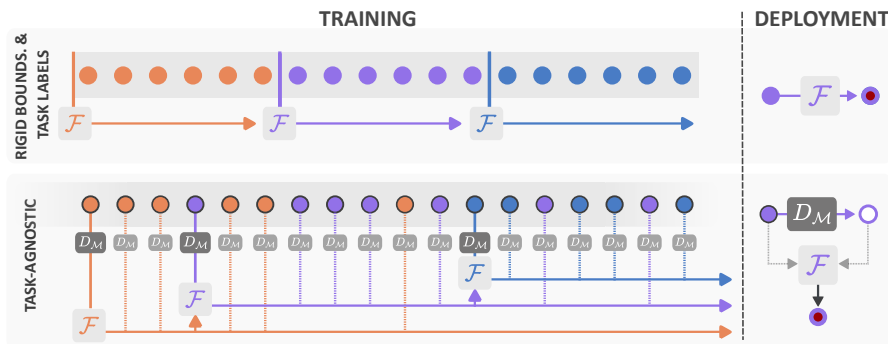
persistent memory, each occupies less than 0.2 GB and the continual OOD detection mechanism ensures that this number remains low. Unlike other methods, *ODEx* requires the same GPU memory and training time as regular sequential learning. Our contributions include:

1. proposing a task-agnostic continual learning solution suitable for a wide array of deep learning architectures, and
2. introducing a continual OOD detection mechanism that does not require access to early data for estimating the distance to the training distribution.

We explore the problem of hippocampus segmentation in T1-weighted MRIs, which is crucial for the diagnosis and treatment of neuropsychiatric disorders but highly sensitive to distribution shifts [25], for two non-stationary environments. Our results show that *ODEx* outperforms state-of-the-art approaches while adhering to desirable properties for continual learning.

## 2 Methodology

We start by defining our problem formulation of task-agnostic continual learning. We then introduce *ODEx*, visualized in Fig. 2 (bottom). During training, we accumulate the mean and covariance of batch normalization layers and detect domain shifts with the Mahalanobis distance. When a domain shift occurs, a new model is initialized with the most appropriate parameters and added to the model pool. During inference, we extract predictions with the best model state.



**Fig. 2.** Top: continual setting with rigid boundaries and task labels. Expansion methods create new parameters at each task boundary. Bottom: the task-agnostic *ODEx* method initializes a new set of parameters when a domain shift is detected.

**Task-agnostic continual learning:** In continual learning settings, model  $\mathcal{F}_\theta : x \rightarrow \hat{y}$  is trained with data samples from an array of  $N_t$  different *tasks* or data distributions  $\{\mathcal{T}_1 \dots \mathcal{T}_{N_t}\}$ , each found at the  $i_{th}$  *stage*  $t_i$ . The model should be

deployable after finishing the first stage, and evolve over time. For segmentation, each instance has the form  $(x, y, j)$ , where  $x$  is an image and  $y$  the segmentation mask. Additionally,  $j$  denotes the *task label*, i.e. that  $(x, y) \sim \mathcal{T}_j$ . The goal is to find parameters  $\theta$  that minimize the loss  $\mathcal{L}$  over all seen tasks  $\{\mathcal{T}_i\}_{i \leq N_t}$  (Eq. 1).

$$\arg \min_{\theta} \sum_{j=1}^{N_t} \mathbb{E}_{(x,y) \sim \mathcal{T}_j} [\mathcal{L}(\mathcal{F}_{\theta}(x), y)] \quad (1)$$

The objective cannot be optimized directly, as at any training stage  $t_j$  only data from  $\mathcal{T}_j$  is available. The main challenge consists of ensuring enough *rigidity* during training to obtain good performance on  $(x, y) \sim \{\mathcal{T}_i\}_{i < j}$  and enough *plasticity* to learn from present and future data  $(x, y) \sim \{\mathcal{T}_i\}_{i \geq j}$ .

*Expansion-based methods* approach this by keeping *task-dependent* parameters  $\{\theta_1 \dots \theta_{N_t}\}$ , which in their simplest form comprise the entire model, and perform inference on  $(x, y, j)$  with the respective  $\mathcal{F}_{\theta_j}$  (see Fig. 2, above). In *task-agnostic scenarios*, task labels  $j$  are unknown and may not even be clearly defined. The goal is to learn a set of parameters  $\Theta = \{\theta_1 \dots \theta_{|\Theta|}\}$  and an inference function  $\mathcal{J} : x \rightarrow \theta$  that selects the best parameters during testing (Eq. 2). In the absence of rigid task boundaries, the size of the model pool  $|\Theta|$  is unknown. Task-agnostic settings thus signify three additional challenges: (1) detecting when domain shifts occur, (2) keeping  $|\Theta|$  low and (3) choosing the best parameters during testing. In the following, we outline how we approach these.

$$\arg \min_{\Theta} \sum_{j=1}^{N_t} \mathbb{E}_{(x,y) \sim \mathcal{T}_j} [\mathcal{L}(\mathcal{F}_{\mathcal{J}(x)}(x), y)] \quad (2)$$

**Detecting domain shifts:** During training, we extract features  $z$  from the first set of *Batch Normalization* layers  $BN_1$ . These normalize inputs and thus contain domain-pertinent information which has been found to play a key role in detecting interference during sequential learning [13]. We estimate a multivariate Gaussian  $\mathcal{N}_i(\mu_i, \Sigma_i)$  at the end of training stage  $t_i$  as:

$$z_k \leftarrow BN_1(x_k); \quad \mu_i \leftarrow \frac{1}{N} \sum_{k=1}^N z_k; \quad \Sigma_i \leftarrow \frac{1}{N} \sum_{k=1}^N (z_k - \mu_i)(z_k - \mu_i)^T \quad (3)$$

Inspired by previous research on OOD detection for semantic segmentation [9], we detect data shifts by calculating the *Mahalanobis distance*  $D_{\mathcal{M}}(z; \mu, \Sigma)$  to the training distribution. In contrast to other methods for assessing similarity, such as the *Gram distance* popular in rehearsal-based continual learning [21,22], the Mahalanobis distance requires storing only  $\mu$  and  $\Sigma$ .

As we cannot revisit data from previous stages, we cannot estimate  $\mathcal{N}$  with all data used to train the model. In a situation with slowly shifting data distributions, if we were to only consider the  $\mu$  and  $\Sigma$  of the last training batch, then we may never detect a sufficiently large distance signaling the need to expand the model pool. We therefore store  $\mu_i$  and  $\Sigma_i$  at the end of each training stage  $t_i$

and add this to the *history*  $\mathcal{B}_i$  of the model which contains information from all pertinent training stages. At stage  $t_{i+1}$ , parameters  $\hat{\theta}$  are selected that minimize the summed distance of the present training data to the history of  $\hat{\theta}$  (Eq. 4).

$$D_{\mathcal{M}}(z; i) : \min_{\theta_j \in \Theta_i} \sum_{(\mu_j, \Sigma_j) \in \mathcal{B}_j} D_{\mathcal{M}_j}(z; \mu_j, \Sigma_j) \quad (4)$$

**Managing the model pool:** When data arrives for a new stage  $t_i$ , the distance  $D_{\mathcal{M}}(z; i)$  is calculated and the best model  $\hat{\theta}$  is selected. If  $D_{\mathcal{M}}(z; i) < \xi$  (case 1), then  $\hat{\theta}$  is updated with the current data. Afterwards,  $\mu_i$  and  $\Sigma_i$  are calculated and added to the model history  $\hat{\mathcal{B}}$ . If instead  $D_{\mathcal{M}}(z; i) \geq \xi$  (case 2), a domain shift is detected and a new model  $\theta_i$  is initialized with the parameters of  $\hat{\theta}$ . After a domain shift, the size of the model pool  $|\Theta|$  grows by 1. The history of the new model  $\mathcal{B}_i$  is initialized with  $\hat{\mathcal{B}}$ , so the history of each model contains information pertaining to all data distributions used to train it. Following previous research [9] we normalize the distances between the minimum and doubled maximum in-distribution values, and set  $\xi = 2\mu$ .

Continuing to train older models instead of initializing a new one for each stage has two advantages: (1) the model pool does not grow linearly with the length of the data stream, which would be prohibiting for deployment over long time periods and (2) models can benefit from further training when the data distributions are compatible, potentially allowing positive backwards transfer.

**Performing inference:** Inference proceeds as illustrated in Fig. 2 (right). For each image, the summed Mahalanobis distance of the test image to each set of parameters  $\theta \in \Theta$  is calculated. Again, the best model  $\hat{\theta}$  is selected and, in this case, directly used to extract a segmentation mask  $\mathcal{F}_{\hat{\theta}}(x) = \hat{y}$ .

### 3 Experimental Setup

We briefly outline how we build our data base of tasks with smooth distribution shifts from publicly available datasets and report relevant aspects of our experimental setup. For further implementation details, we refer the reader to the supplementary material and our code found under <https://github.com/MECLabTUDA/Lifelong-nnUNet>.

**Data:** We look at two different scenarios of data streams with slowly shifting distributions for segmentation of the entire hippocampus (head, body and tail) in T1-weighted MRIs. The first is constructed from three public datasets: *HarP* [3] contains 135 healthy and Alzheimer’s disease patients, *Dryad* [15] has 25 healthy adult subjects and *Decathlon* [26] contains 130 healthy and schizophrenia patients. We slowly shift the distribution of cases from each source as illustrated in Appendix A. We refer to this scenario as **shifting source**. For the second scenario, henceforth referred to as **transformed**, we slowly modify the *Decathlon* data using the *TorchIO* library [20]. We apply intensity rescaling up to a contrast stretching of (0.1, 0.9) and affine transformations of up to a (0.8, 1.2) scaling range, 15 degrees rotation and 5 mm translation.

**Table 1.** Performance of the joint training upper bound (first row), sequential learning and six continual learning strategies on the two hippocampus segmentation scenarios.

Method	Shifting source			Transformed		
	Dice $\uparrow$	BWT $\uparrow$	FWT $\uparrow$	Dice $\uparrow$	BWT $\uparrow$	FWT $\uparrow$
Joint	.89 $\pm$ .01			.90 $\pm$ .01		
Seq.	.57 $\pm$ .32	-.19 $\pm$ .12	.14 $\pm$ .09	.87 $\pm$ .03	-.02 $\pm$ .02	.09 $\pm$ .05
EWC	.78 $\pm$ .08	.02 $\pm$ .03	.08 $\pm$ .08	.79 $\pm$ .10	.01 $\pm$ .01	.04 $\pm$ .02
MiB	.67 $\pm$ .24	-.10 $\pm$ .07	.14 $\pm$ .10	.87 $\pm$ .04	-.02 $\pm$ .02	.07 $\pm$ .04
RW	.61 $\pm$ .28	-.15 $\pm$ .10	.14 $\pm$ .10	.87 $\pm$ .03	-.03 $\pm$ .03	.09 $\pm$ .05
PLOP	.57 $\pm$ .32	-.22 $\pm$ .14	.13 $\pm$ .09	.86 $\pm$ .02	-.02 $\pm$ .02	.10 $\pm$ .06
LwF	.51 $\pm$ .35	-.23 $\pm$ .13	.10 $\pm$ .07	.86 $\pm$ .04	-.04 $\pm$ .04	.10 $\pm$ .06
ODEx (ours)	.87 $\pm$ .04	-.03 $\pm$ .02	.14 $\pm$ .09	.89 $\pm$ .01	-.01 $\pm$ .01	.09 $\pm$ .05

**Network architecture and training:** We use a full-resolution *nnUNet* [11] model for all experiments, with the architecture and training settings selected for the first training stage of each data stream. We perform 200 epochs for each stage, with a loss of *Dice* and *Binary Cross Entropy* weighted equally. All experiments were carried out on a *Nvidia Tesla T4* GPU (16 GB).

**Metrics:** We report the average Dice on test data from all tasks  $\{\mathcal{T}_i\}_{i \leq N_t}$  as well as backwards (BWT) and forwards (FWT) transferability [7,10]. BWT is the *inverse forgetting* and displays to what extent the performance on test samples  $(x, y) \sim \mathcal{T}_i$  deteriorates with further training in stages  $\{t_i\}_{i > N_t}$ . FWT instead measures what impact training on each stage  $\{t_i\}_{i \leq N_t}$  has on test data  $(x, y) \sim \mathcal{T}_i$ . Methods that prevent forgetting show high, realistically close to 0, BWT. FWT is high if enough plasticity is maintained to acquire new knowledge. For both metrics, we report the average over test data from all tasks.

**Baselines:** In Sec. 4.1, we compare our approach against sequential training and five popular continual learning approaches: Elastic Weight Consolidation (EWC) [14], Modelling the Background (MiB) [4], Riemannian Walk (RW) [5], PLOP [8] and Learning without Forgetting (LwF) [17]. We also report the upper bound of joint training. In most cases, we use the hyperparameters suggested in the corresponding publications or code bases (for more details see Appendix B). For MiB, we reduce the *lkd* to prevent loss explosion. In Sec. 4.2 we perform an ablation study and compare the use of the Mahalanobis distance to other methods proposed within task-agnostic learning, namely using the Gram matrix [21] and detecting domain shifts through a fall in training performance [6].

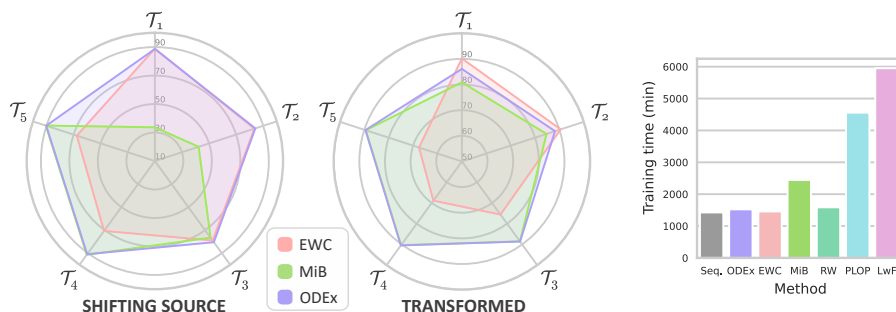
## 4 Results

We first compare *ODEx* to state-of-the-art continual learning approaches in Sec. 4.1. Afterwards, we take a closer look at the cumulative Mahalanobis distance for identifying domain shifts and selecting the best parameters (Sec. 4.2).

#### 4.1 Continual learning performance

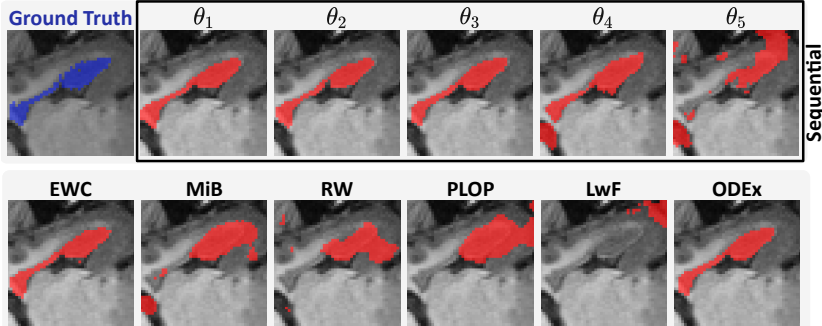
We compare our proposed approach *ODEx* to five continual learning methods in Tab. 1. The first row shows the upper bound of training a model statically with all training data. *Sequential* results show the deterioration of the performance in earlier tasks as training is carried out, and the following rows display how five continual learning strategies alleviate this. From these, only *EWC* maintains performance on earlier tasks, but at the cost of losing model plasticity and being unable to acquire new knowledge. *ODEx* instead reaches a high FWT showing effective learning on later tasks while still performing well on data from the first training stages. This behavior is further illustrated in Fig. 3 (left), where the per-task performance is plotted for *EWC*, which successfully retains old knowledge, *MiB*, which reaches a high Dice on later tasks, and *ODEx* that performs well on data from all stages. This is particularly clear for the more difficult *shifting source* case, but a Wilcoxon one-sided signed-rank test affirms that *ODEx* significantly outperforms all other approaches in terms of Dice score for both scenarios.

As for resource utilization, *ODEx* requires no more GPU memory than sequential training, as we update one model at a time. The estimation of  $\Sigma$  and the calculation of  $D_{\mathcal{M}}$  can be carried out in the CPU given the low resolution of  $z$ . Fig. 3 (right) shows that *ODEx* takes only marginally longer than training without any method for forgetting prevention. Though several models are stored (two for *shifting source* and four for *transformed*, see Tab. 2) each weights less than 200 MB, being far from a limiting factor in practice.



**Fig. 3.** Left: Per-task Dice. EWC and MiB are at opposite ends of the plasticity/rigidity spectrum, whereas ODEx allows for further training without compromising performance on previous tasks. Right: training times for the shifting source scenario.

Fig. 4 qualitatively shows in the upper row the sequential deterioration of the segmentation for a test subject  $(x, y) \sim \mathcal{T}_1$ . The lower row displays the segmentation masks produced by each continual learning method. Though the head is mostly segmented well by several methods, only *EWC* and *ODEx* properly segment the body and tail and maintain the integrity of the shape.



**Fig. 4.** Crops with overlaid segmentations for axial slice 25 of a subject from  $\mathcal{T}_1$  (shifting source). Top: ground truth (blue) and performance deterioration with regular SGD. Bottom: six continual learning methods, after finishing training on last stage.

## 4.2 Ablation study

In Tab. 2 we compare our strategy for detecting when to grow the model pool to previous work in the field of task-agnostic learning. The performance of all methods is very similar for the easier *transformed* scenario, but we see clear differences in *shifting source*. We first explore two versions of *ODEx* that use our proposed strategy for selecting the best model but detect domain shifts in a different fashion. *ODEx*  $-\infty \xi$  creates a new model for every stage. The lower Dice suggests that the models suffer from the lack of training data, and  $|\Theta|$  grows linearly with the number of training stages. *DiceEx* initializes a new model when the training Dice falls more than 10%, which results in higher forgetting. *ODEx*  $-\mathcal{B}$  shows the situation where we do not keep a history for the training distributions of previous stages and only calculate the distance to the last stage. For this version, no new model is initialized for *shifting source* and the single available model significantly forgets previous knowledge. Finally, we test the use of the Gram distance instead of Mahalanobis for both training and testing, and find that it does not properly detect distribution shifts for *shifting source*.

**Table 2.** Performance of different strategies for detecting domain boundaries and/or selecting a model state during inference.

Method	Shifting source				Transformed			
	Dice $\uparrow$	BWT $\uparrow$	FWT $\uparrow$	$ \Theta  \downarrow$	Dice $\uparrow$	BWT $\uparrow$	FWT $\uparrow$	$ \Theta  \downarrow$
ODEx (ours)	<b>.87</b> $\pm$ .04	-.03 $\pm$ .02	<b>.14</b> $\pm$ .09	2	.89 $\pm$ .01	-.01 $\pm$ .01	<b>.09</b> $\pm$ .05	4
ODEx $-\infty \xi$	.83 $\pm$ .04	<b>.00</b> $\pm$ .00	.11 $\pm$ .09	5	.89 $\pm$ .01	<b>.00</b> $\pm$ .00	<b>.09</b> $\pm$ .05	5
DiceEx [6]	.84 $\pm$ .08	-.07 $\pm$ .03	<b>.14</b> $\pm$ .10	2	.89 $\pm$ .02	-.01 $\pm$ .01	<b>.09</b> $\pm$ .05	<b>2</b>
ODEx $-\mathcal{B}$ [9]	.57 $\pm$ .32	-.19 $\pm$ .12	<b>.14</b> $\pm$ .09	1	.89 $\pm$ .01	<b>.00</b> $\pm$ .00	<b>.09</b> $\pm$ .05	3
Gram [21]	.57 $\pm$ .32	-.19 $\pm$ .12	<b>.14</b> $\pm$ .09	1	<b>.90</b> $\pm$ .01	<b>.00</b> $\pm$ .00	<b>.09</b> $\pm$ .05	3

## 5 Conclusion

We introduce *ODEx*, an expansion-based continual learning strategy suitable for real clinical environments with smooth acquisition and population shifts. We evaluate our approach on two hippocampus segmentation scenarios and show that it outperforms state-of-the-art methods by maintaining good performance on data from early stages without compromising model plasticity. *ODEx* requires only marginally higher training times than regular sequential learning, and the same amount of GPU memory. While additional persistent storage is needed to store different sets of parameters, the OOD detection strategy keeps this number low. Each explored scenario required less than 0.8 GB, rendering this limitation insignificant in practice. Future work should explore whether it suffices to maintain only a subset of domain-specific parameters, such as the last decoder blocks or batch normalization layers. By releasing our code and models, we hope to boost continual learning research in task-agnostic medical settings.

## References

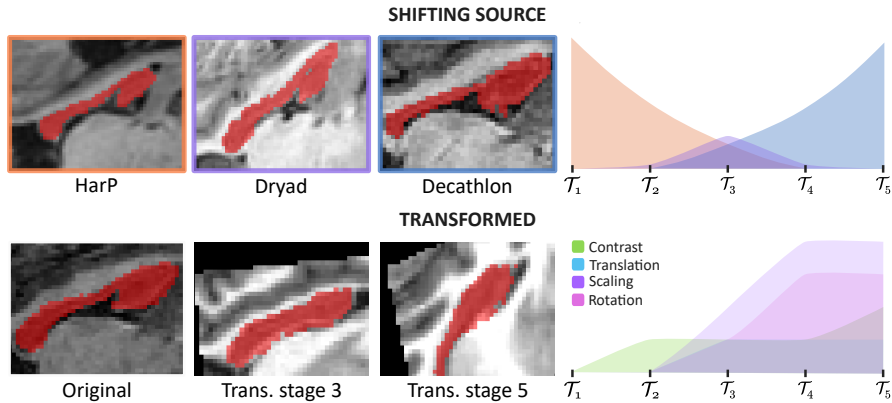
1. Aljundi, R., Belilovsky, E., Tuytelaars, T., Charlin, L., Caccia, M., Lin, M., Page-Caccia, L.: Online continual learning with maximal interfered retrieval. *NeurIPS* **32** (2019)
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. *NeurIPS* **32** (2019)
3. Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani, M., et al.: Training labels for hippocampal segmentation based on the eadc-adni harmonized hippocampal protocol. *Alzheimer’s & Dementia* **11**(2), 175–183 (2015)
4. Cermelli, F., Mancini, M., Bulò, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: *CVPR*. pp. 9233–9242 (2020)
5. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: *ECCV*. pp. 532–547 (2018)
6. Chen, H.J., Cheng, A.C., Juan, D.C., Wei, W., Sun, M.: Mitigating forgetting in online continual learning via instance-aware parameterization. *NeurIPS* **33**, 17466–17477 (2020)
7. Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
8. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: *CVPR*. pp. 4040–4050 (2021)
9. Gonzalez, C., Gotkowski, K., Bucher, A., Fischbach, R., Kaltenborn, I., Mukhopadhyay, A.: Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 304–314. Springer (2021)
10. Hadsell, R., Rao, D., Rusu, A.A., Pascanu, R.: Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences* **24**(12), 1028–1040 (2020)

11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
12. Jin, X., Sadhu, A., Du, J., Ren, X.: Gradient-based editing of memory examples for online task-free continual learning. *NeurIPS* **34** (2021)
13. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain mr segmentation across scanners and protocols. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 476–484. Springer (2018)
14. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* **114**(13), 3521–3526 (2017)
15. Kulaga-Yoskovitz, J., Bernhardt, B.C., Hong, S.J., Mansi, T., Liang, K.E., Van Der Kouwe, A.J., Smallwood, J., Bernasconi, A., Bernasconi, N.: Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset. *Scientific Data* **2**(1), 1–9 (2015)
16. Lao, Q., Jiang, X., Havaei, M., Bengio, Y.: Continuous domain adaptation with variational domain-agnostic feature replay. *arXiv preprint arXiv:2003.04382* (2020)
17. Li, Z., Hoiem, D.: Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence* **40**(12), 2935–2947 (2017)
18. Memmel, M., Gonzalez, C., Mukhopadhyay, A.: Adversarial continual learning for multi-domain hippocampal segmentation. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 35–45. Springer (2021)
19. Özgün, S., Rickmann, A.M., Roy, A.G., Wachinger, C.: Importance driven continual learning for segmentation across domains. In: *International Workshop on Machine Learning in Medical Imaging*. pp. 423–433. Springer (2020)
20. Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine* p. 106236 (2021). <https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106236>, <https://www.sciencedirect.com/science/article/pii/S0169260721003102>
21. Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G.: Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nature Communications* **12**(1), 1–12 (2021)
22. Perkonigg, M., Hofmanninger, J., Langs, G.: Continual active learning for efficient adaptation of machine learning models to changing image acquisition. In: *International Conference on Information Processing in Medical Imaging*. pp. 649–660. Springer (2021)
23. Prabhu, A., Torr, P.H., Dokania, P.K.: Gdumb: A simple approach that questions our progress in continual learning. In: *European conference on computer vision*. pp. 524–540. Springer (2020)
24. Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y.W., Hadsell, R.: Continual unsupervised representation learning. *NeurIPS* **32** (2019)
25. Sanner, A., Gonzalez, C., Mukhopadhyay, A.: How reliable are out-of-distribution generalization methods for medical image segmentation? In: *DAGM German Conference on Pattern Recognition*. pp. 604–617. Springer (2021)
26. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B.H., Ronneberger,



- O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W.R., McHugo, M., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR* **abs/1902.09063** (2019)
27. Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z., Mahapatra, D.: Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 226–238. Springer (2021)
  28. Venkataramani, R., Ravishankar, H., Anamandra, S.: Towards continuous domain adaptation for medical imaging. In: *IEEE 16th ISBI*. pp. 443–446. IEEE (2019)
  29. Vokinger, K.N., Gasser, U.: Regulating ai in medicine in the united states and europe. *Nature machine intelligence* **3**(9), 738–739 (2021)
  30. Zhang, J., Gu, R., Wang, G., Gu, L.: Comprehensive importance-based selective regularization for continual segmentation across multiple sites. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 389–399. Springer (2021)
  31. Zheng, E., Yu, Q., Li, R., Shi, P., Haake, A.: A continual learning framework for uncertainty-aware interactive image segmentation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 6030–6038 (2021)

## A Data scenarios



**Fig. 1.** The two scenarios of data streams with distribution shifts explored in this work. Top: number of cases from three datasets is slowly shifted. Bottom: the *Decathlon* dataset is artificially transformed. We used the first 80/20 split generated by the nnUNet framework for *HarP*, *Dryad* and *Decathlon* and ensured that test cases remained as such across both scenarios.

## B Architecture and training parameters

**Table 1.** Hyperparameters for training continual learning methods. The settings specified in the first row were used for all experiments.

Method	Setting
All	optimizer = SGD, lr = 0.01, weight decay = $3e - 5$ , momentum = .99, nr. blocks = 4 for <i>shifting source</i> , nr. blocks = 3 for <i>transformed</i>
EWC	$\lambda = 0.4$
MiB	$\alpha = 0.9$ , lkd = 1 for <i>shifting source</i> , lkd = 0.1 for <i>transformed</i>
RW	$\alpha = 0.9$ , $\lambda = 0.4$ , update after = 10
PLOP	$\lambda = 0.01$ , scales = 3, resampling to (48, 48, 48) for <i>transformed</i> , no resampling for <i>shifting source</i>
LwF	$T = 2$

## C Calculation of evaluation metrics

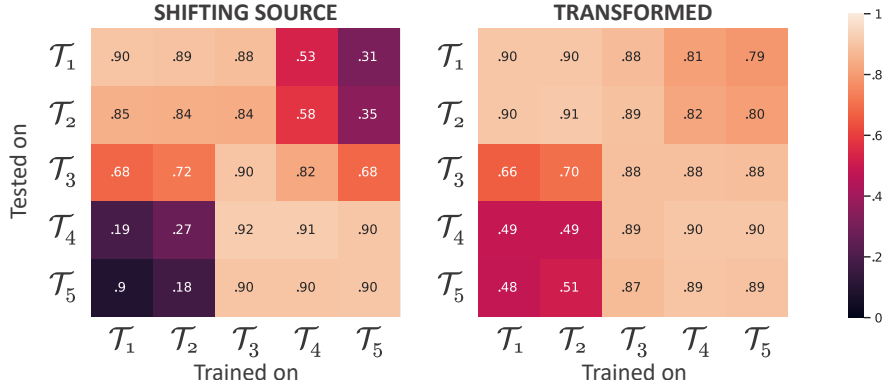
Considering  $\mathcal{F}_i(x) = \hat{y}_i$  as the prediction made at stage  $t_i$ , backwards transfer (BWT) is the change in performance after training with each subsequent task  $\{\mathcal{T}\}_{j>i}$ , averaged over the number of samples in  $\mathcal{T}_i$  and the number of tasks (Eq. 1). BWT is not defined for the last task  $\mathcal{T}_{N_t}$ , as  $\{t_j\}_{j>N_t} = \emptyset$ .

$$BWT = \frac{1}{N_t} \sum_{i=1}^{N_t} \left[ \frac{1}{|\{t_j\}_{j>i}|} \sum_{j>i} \left[ \frac{1}{|\mathcal{T}_i|} \sum_{k=1}^{|\mathcal{T}_i|} \text{Dice}(\mathcal{F}_j(x_k), y_k) - \text{Dice}(\mathcal{F}_i(x_k), y_k) \right) \right] \right] \quad (1)$$

Forwards transfer (FWT) is, for each task  $\mathcal{T}_i$ , the change in performance in each stage before and up to  $t_i$ , averaged over the number of samples and tasks. FWT is not defined for the first task  $\mathcal{T}_1$ , as  $\{t_j\}_{j<1} = \emptyset$ .

$$FWT = \frac{1}{N_t} \sum_{i=1}^{N_t} \left[ \frac{1}{|\{t_j\}_{j\leq i}|} \sum_{j\leq i} \left[ \frac{1}{|\mathcal{T}_i|} \sum_{k=1}^{|\mathcal{T}_i|} \text{Dice}(\mathcal{F}_j(x_k), y_k) - \text{Dice}(\mathcal{F}_{j-1}(x_k), y_k) \right) \right] \right] \quad (2)$$

## D Static learning results



**Fig. 2.** Base transferability in terms of Dice score of training separate models statically with each task on test data from each task.

---

## 8.2. Conclusions and outlook

---

My practical experience working with an array of continual learning strategies shows that expansion-based solutions are the most practical in terms of performance (as they preserve previous knowledge without affecting plasticity), computational overhead (as they merely require additional *persistent* memory, which is rarely a constraint) and flexibility (as they can augment any existing architecture and do not disturb the training process).

One central reason ensembles are so popular is that they do not require making additional design decisions of data preparation, architecture, or training procedure. Just as deep ensembles have demonstrated to be a simple strategy that leads to performance increases and usable uncertainty estimates, expansion-based continual learning approaches have the potential to easily transfer DNNs to the clinical open world. And just as ensembles can be made more efficient by sharing certain weights, we can take the same approach with continual learning. Parameter-sharing solutions (such as multi-head architectures) have a long-standing tradition in continual learning and could potentially allow for positive backward transfer by sharing parameters of modules that solve similar aspects of the problem.

As mentioned previously, one direction I am optimistic about and exploring at the moment is to *combine* the predictions of multi-model solutions as if they were an ensemble. That is to say, selecting a *set* of models instead of just one for both (1) updating their parameters during training and (2) carrying out inference. Perhaps, depending on the model, only certain layers should be used to construct a sub-ensemble that is the most suitable for making each prediction.

One central takeaway I would like to leave from my work is that solving the continual learning problem cannot be separated from addressing distribution shifts and detecting changes in the domain. Purely algorithmic solutions that do not consider *what* changes transpire in the data and proactively react to them will rarely learn expressive representations that can leverage all training samples and reach top performance across domains.

---

## 9. Learning Meaningful Representations

---

Our goal, then, is not merely to prevent catastrophic forgetting while allowing the model to learn as easily as if it was being fine-tuned. We ought to go beyond that and instead search for *positive backward transfer*. That is, to accumulate knowledge from all the data we see during the training process. We hope to hereby *improve* across *all* domains and match the performance of the statically trained upper bound.

If we wish to achieve this with one single model, then we need to learn *expressive features* that encode as much relevant knowledge as possible. Similarly, we want to disregard information that is not relevant and instead leads models to learn spurious correlations. If we successfully build such a feature space, there will be no interference caused by non-relevant image characteristics. This would also improve generalizability to new domains.

This research direction is tightly connected to the OOD generalization and domain adaptation strategies reviewed in Chapter 6. Unfortunately, when we pursue this avenue, we come across the same challenges we faced then: greater computational requirements, changes to the architecture or training objective that limit flexibility, and ultimately disappointing performance. Considering this, I do not believe this to be the most practical strategy for medical imaging *at the moment*, though it is a fascinating field.

Instead of augmenting models with new components for continual learning, it is also worth quantifying whether specific layers or practices in DNN design would naturally lead to less forgetting. For instance, visual transformers, which have gained enormous popularity in recent years, display certain properties that could be advantageous for continual learning. I look into this possibility in Section 9.2.

---

### 9.1. Building an expressive latent space

---

Similarly to the domain adaptation methods we explored in past sections, we can use *generative adversarial networks* (Goodfellow et al., 2020) to build representations that disregard certain data characteristics, such as image properties that result from the acquisition process. Several works look into this strategy for computer vision problems (Ebrahimi et al., 2020; Michieli and Zanuttigh, 2021), and Elskhawy et al. (2020) propose a method with an adversarial component for an incremental class learning scenario on chest CTs. The approach outperforms regular fine-tuning and *LwF*, but maintains task-specific heads and does not show an improvement against individual model states.

We similarly present a GAN-based approach for hippocampus segmentation (Mommel et al., 2021). The *Adversarial Continual Segmenter (ACS)* assumes that we have access to at least two different domains in the initial training phase and adversarially disentangles the domain information from the content. By removing domain-identifying information from this new data, the last few layers can be fine-tuned while preserving more knowledge. However, the proposed architecture is rather particular and involves several discriminator and generator modules in addition to the segmentation section. It is, therefore, clearly not

---

a flexible solution that can be applied to any problem and model. In addition, the empirical results show a decrease in forgetting but a significant gap to state-of-the-art segmentation performance for the task.

A different – though related – approach is to translate features from new incoming domains to the source domain to prevent interference. Ravishankar et al. (2019) explore this strategy for X-Ray pneumothorax and ultrasound cardiac view classification, showing an improvement in knowledge preservation though no direct performance gains.

---

## 9.2. Continual learning in transformer architectures

---

Transformer architectures (Vaswani et al., 2017), and specifically vision transformers (ViTs) for computer vision (Dosovitskiy et al., 2021), are becoming hugely popular. Some recent work shows that such components, which learn the relation between sequences of image patches, could be helpful for medical image segmentation (Karimi et al., 2021; Hatamizadeh et al., 2022).

In a recent article (Ranem et al., 2022), we examine whether transformer-based segmentation architectures are less affected by the problem of catastrophic forgetting given how they combine sequentially arranged inputs with global knowledge. A major contribution of our work is the implementation of a vision transformer option in *Lifelong nnU-Net*, our framework extending the *nnU-Net* pipeline for continual learning. Specifically, we design a model that places the Vision Transformer (ViT) block between the encoding and decoding layers. Our results indicate that this has a minimal effect on knowledge preservation. In addition, augmenting transformer architectures with continual learning methods can be challenging. In particular, we find that regularizing ViT layers with *EWC* can lead to decreased performance.

Given the rise in popularity of ViTs for medical segmentation, our study provides interesting insight into how these models react when trained in a sequential fashion. Nevertheless, from the continual learning perspective, this type of architecture does not display better properties than a regular U-Net.

---

## 9.3. Conclusions and outlook

---

I still believe in the overarching goal of learning expressive, semantically sound feature representations that capture all the diagnostically-relevant content of the training images. Such a latent space could be sequentially updated with less interference. Nevertheless, our empirical experience shows that several factors speak against the practicality of this research direction for continual learning (at least at the moment). For starters, generative modeling and/or feature disentanglement comprise a significant computational overhead. It can also be challenging to reach convergence, as multiple loss terms need to be appropriately weighted to reach training equilibrium. Further, purely domain-agnostic features often do not result in the expected performance on the downstream target task.

New architectures and training mechanisms are being proposed for image classification and segmentation, which may be more prone to knowledge preservation. Though I would not recommend adding a particular component to a model for the purpose of continual learning, by routinely evaluating new architectures in a continual fashion – the way we do for robustness – we could identify new layers or mechanisms that are naturally better suited to dynamic environments.

---

## **Part III.**

# **Towards Lifelong Learning in the Clinical Workflow**

---

## 10. Practical Challenges Hindering Lifelong Learning

---

We have looked in length at the problem of data drift in medical imaging and discussed technical methods to learn continuously. Unfortunately, we are far from bringing these advances to clinical practice. In this part of the thesis, I will take a more holistic view at the practical challenges hindering the use of systems that adapt to changing clinical environments. In the past three years, I have helped overcome these difficulties by establishing benchmarking standards and evaluation best practices, releasing open-source projects that we thought would be helpful to the community, and maintaining a dialogue with different stakeholders on how *they* approach the problem of data drift. I presented much of the work shown in this section at the radiological venues *EuSoMII*, *RSNA* and *ECR*.

The first obstacle we must overcome is the need for appropriate reporting standards. Benchmark datasets commonly used in continual learning research, such as *split* and *permuted MNIST*, are overly simplistic and unsuitable for medical imaging. Lacking a unified framework where researchers can evaluate methods across anatomies and imaging modalities, it is challenging to compare different works, and the barrier for developing new techniques is high. We approach this for image segmentation with our *Lifelong nnU-Net* project (Chapter 11), where we augment the state-of-the-art *nnU-Net* pipeline with continual learning methods and metrics and report results for three anatomies using openly-available datasets.

Another key element for developing adequate decision support systems lies in looking at the larger clinical workflow where several DL models work alongside healthcare professionals. In Chapter 12, we describe a scenario where a radiologist receives a second opinion on the localization of pulmonary emboli. We show how quality assurance mechanisms, interpretability techniques, and an active learning approach can all contribute to effective collaboration with the radiologist.

I hope to have painted a convincing picture of how DNNs can be used safely and in close interaction with clinicians, leveraging user input to maintain and even *increase* performance over time. Nevertheless, companies are reluctant to develop products that learn continuously given the difficulty that these be cleared for commercial use. Instead, they invest significant resources in assembling large and heterogeneous databases with the hope that the trained models will be robust as long as possible. Though many products in the market today *collect* user feedback in the form of corrected diagnosis, they do not use this data until a new product release months or even years later. When queried about this strategy – and continual learning in general – they maintain that current regulations strictly prohibit frequent model updates and this will remain so for the foreseeable future.

While it is true that current regulatory frameworks do not allow for continual learning, there are initiatives from both the US Food and Drug Administration (FDA) and European Commission indicating that this could change sooner than many stakeholders expect. In particular, the interest of regulatory bodies in maintaining safety and efficiency throughout the entire lifecycle of medical devices and the focus on active quality assurance suggest that regulatory changes are fast approaching. In Chapter 13, I summarize regulations currently in force for medical software in the US and European Union, as well as official documents that hint at how the clearance process may soon be adapted for lifelong learning.



---

## 11. The Need for Unified Evaluation Standards

---

Despite increased interest in continual learning from the medical imaging community (for instance, *continual learning* is now a paper category in the MICCAI conference), the number of publications does not reflect how relevant the topic is for maintaining consistent performance of medical software. Particularly for image segmentation, regularization-based approaches are still regarded as a satisfactory solution, and most studies only cover one anatomy (Baweja et al., 2018; van Garderen et al., 2019; Özgün et al., 2020; Patra and Noble, 2020; Zhang et al., 2021).

We are even further from a situation where papers on various topics include evaluations in continual environments the way they do on *external*/OOD data. Quantifying how well DNNs designed for diverse problems react to post-deployment adaptation would encourage the release of lifelong learning products much more than dedicated continual learning research.

One major reason behind this is the lack of open-source continual learning frameworks and unified evaluation standards. I further believe that only models that reach the best performance in the target task have a real chance to come into use. Modifications to the model architecture which cause performance degradation will be avoided even if this fulfills a secondary goal such as OOD detection or forgetting prevention. Therefore, benchmark evaluations for continual learning should be performed in state-of-the-art frameworks.

---

### 11.1. The paper: Lifelong nnU-Net for standardized medical continual learning

---

We attempt to alleviate these issues with *Lifelong nnU-Net*, an open-source project where we extend the popular *nnU-Net* library, which sits at the top of numerous medical segmentation challenges (Isensee et al., 2021), with continual learning capabilities. The functionality includes multiple continual learning methods and all the required logic to monitor performance across datasets over time. The project – hosted in [github.com/MECLabTUDA/Lifelong-nnUNet](https://github.com/MECLabTUDA/Lifelong-nnUNet) – received a good reception from the community, currently bearing 72 stars and 9 forks.

Alongside the code, we drafted the manuscript *Lifelong nnU-Net: a framework for standardized medical continual learning* (González et al., 2022c), where we report benchmark results on openly available datasets for three anatomies and define evaluation standards for continual learning in medical imaging. The paper is the result of a close collaboration with Daniel Pinto dos Santos from the University Hospital Cologne and the neuroradiologist Ahmed Othman from the University Medical Center Mainz. From the TU Darmstadt, Amin Ranem and Anirban Mukhopadhyay helped carry out the experiments and draft the text. The manuscript is currently under review, but we already presented an abstract at the *European Society of Medical Imaging Informatics (EuSoMII) Annual Meeting* this past October in Valencia, where I was awarded the *best oral presentation* award. The content was also accepted for oral presentation at the *European Congress of Radiology (ECR)* taking place in March 2023.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**Lifelong nnU-Net: a framework for standardized medical continual learning**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "Lifelong nnU-Net: a framework for standardized medical continual learning" (González et al. 2022) is currently under review. It constitutes a joint work of Camila González, Amin Ranem, Daniel Pinto dos Santos, Ahmed Othman and Anirban Mukhopadhyay.*

*This work was supported by the Bundesministerium für Gesundheit (BMG) with grant [ZMVI1-2520DAT03A].*

*As corresponding and leading author, C. González led the overall research design, literature review and writing process of the paper. The choice of methodological framework and experimental setup were likewise made by C. González. The implementation of the code was performed by C. González and A. Ranem. A. Ranem conducted the experiments. C. González and A. Ranem contributed to the analysis of the data and results. The methodology, results and discussion were mainly written by C. González. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor, who also contributed with continuous feedback during all phases of the paper writing process. D. Pinto dos Santos and A. Othman motivated and reviewed the manuscript from a clinical perspective. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum: 01 / 30 / 2023 01 / 30 / 2023 01 / 30 / 2023 01 / 30 / 2023

Unterschrift:    

Camila González

Amin Ranem

Daniel Pinto dos Santos

Ahmed Othman

Datum: 01 / 30 / 2023

Unterschrift: 

Anirban Mukhopadhyay

# Lifelong nnU-Net: a framework for standardized medical continual learning

Camila González<sup>1,\*</sup>, Amin Ranem<sup>1</sup>, Daniel Pinto dos Santos<sup>2,3</sup>, Ahmed Othman<sup>4</sup>, and Anirban Mukhopadhyay<sup>1</sup>

<sup>1</sup>Technical University of Darmstadt, Karolinenpl. 5, 64289 Darmstadt, Germany

<sup>2</sup>University Hospital Cologne, Kerpener Str. 62, 50937 Cologne, Germany

<sup>3</sup>University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt, Germany

<sup>4</sup>University Medical Center Mainz, Langenbeckstraße 1, 55131 Mainz, Germany

\*camila.gonzalez@gris.tu-darmstadt.de

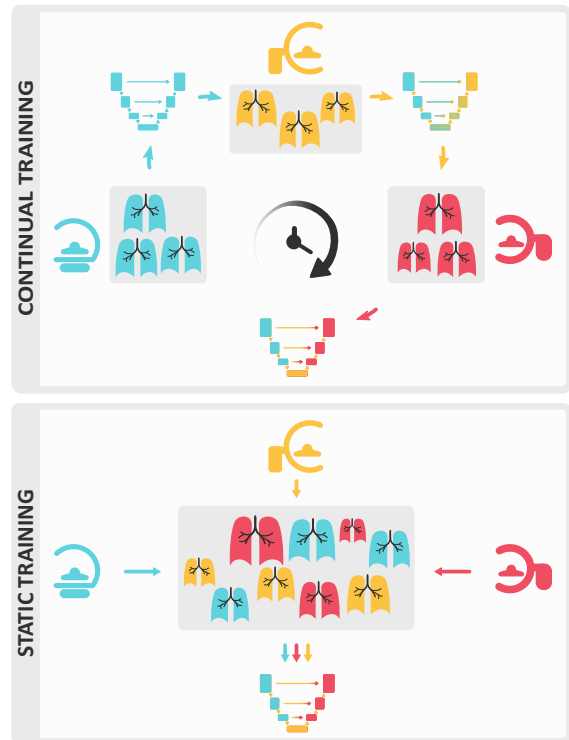
## ABSTRACT

As the enthusiasm surrounding Deep Learning grows, both medical practitioners and regulatory bodies are exploring ways to safely introduce image segmentation in clinical practice. One frontier to overcome when translating promising research into the clinical open world is the shift from *static* to *continual* learning. Continual learning, the practice of training models throughout their lifecycle, is seeing growing interest but is still in its infancy in healthcare. We present *Lifelong nnU-Net*, a standardized framework that places state-of-the-art continual segmentation at the hands of researchers and clinicians. Built on top of the nnU-Net – widely regarded as the best-performing segmenter for medical applications – and equipped with all necessary modules for training and testing models sequentially, we ensure broad applicability and lower the barrier to evaluating new methods in a continual fashion. Our benchmark results across three medical segmentation use cases and five continual learning methods give a comprehensive outlook on the current state of the field and signify a first reproducible benchmark.

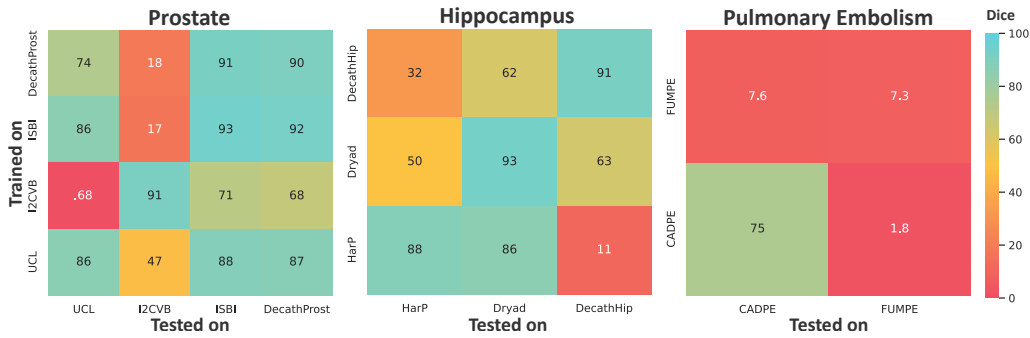
## Introduction

Deep Learning methods for medical use cases continue to be evaluated in a *static* setting, where all available data is shuffled and the model is trained on a subset of in-distribution samples. This stands on the unrealistic assumptions that (a) all training data is available in a central location, and (b) the acquisition conditions do not change over time after clinical deployment<sup>1</sup>. Evaluating in this manner creates a considerable gap between the reported performance of new methods and their usability in practice<sup>2-4</sup>, which hinders the vital deployment of lifelong learning agents in dynamic clinical environments<sup>5</sup>.

*Continual learning* does not neglect the temporal dimension of the data and trains models in a sequential fashion, as illustrated in Figure 1. The goal hereby is to adapt to new environments without losing performance on previously observed training conditions and subject groups. Distributed *federated learning* methods have been explored in multi-clinical settings and also do not require sharing data between institutions<sup>6,7</sup>. However, they neither address *temporal restrictions on data availability* nor provide a framework for agents that continuously adapt to shifting population dynamics. Continual learning in healthcare, which tackles these concerns, is receiving growing enthusiasm<sup>8-11</sup> and regulatory procedures are being actively debated<sup>5,12,13</sup>. Currently, re-approval is required each time a model is adapted during deployment, but there are initiatives from both the FDA and European Commission for a *lifecycle regulatory protocol* that allows the use of continuously adapting algorithms<sup>14</sup>. These pursuits may lead us to the rare situation where the regulatory guidelines are in place while the technology is still in its infancy.



**Figure 1.** In a static setting (bottom), all training data is brought together. Continual settings (top) consider the time of acquisition and train the model sequentially.



**Figure 2.** Performance of models trained independently solely on one dataset. On the (lower left to upper right) diagonal we find the Dice coefficient of evaluating models on the test cases of the dataset used for training. In the remaining cells, we see how these models transfer to other datasets.

Technical literature of continual learning for simpler computer vision tasks is plagued by controversies about the lack of a standardized evaluation setup<sup>15–17</sup>. Recently, the *Avalanche*<sup>18</sup> project has emerged as a solution to this problem for continual classification by providing a unified code base. The field is not as mature for continual *segmentation*, which assigns a label to each pixel in the image and is arguably the primary AI task in the clinical domain. Though more work has been done in recent years<sup>8, 10, 19–23</sup>, it neither (1) builds on top of state-of-the-art segmentation pipelines nor (2) examines how popular methods transfer to image segmentation for multiple open-source benchmarks.

In this work, we present *Lifelong nnU-Net*, a standardized framework for training and evaluating segmentation models in continual settings. We build our code on top of the *nnU-Net*<sup>24</sup> pipeline, which is widely popular and state-of-the-art for medical segmentation tasks, thus ensuring high usability and performance. Our contributions are:

- the introduction of an open-source continual learning framework built on top of the state-of-the-art *nnU-Net*
- a thorough performance and run-time comparison for training sequentially under different settings, and
- open-source implementations for *five continual learning methods*, allowing the fast evaluation of the state-of-the-art and accelerating the development of new approaches.

Our experiments on publicly available data for three different segmentation problems show that:

- none of the explored continual learning methods consistently achieve *positive backward transfer* for segmentation, exhibiting the need for new solutions,
- in accordance with previous research, rehearsal-based methods display the least amount of forgetting while maintaining model plasticity, and
- the practice of maintaining task-specific heads, common in continual learning literature, is only minimally relevant for segmentation.

The goal of *Lifelong nnU-Net* is to ensure high technical standards and reproducible results while the community is translating continual learning to medical image segmentation. By *releasing our code and trained models* for open-source datasets, we establish a benchmark for evaluating future continual learning methods on segmentation models.

## Results

We start this section by examining the results of training models statically with one dataset. Afterward, we explore sequential learning and five popular continual learning strategies: *Rehearsal*, *Elastic Weight Consolidation*<sup>25</sup> (EWC), *Learning without Forgetting*<sup>26</sup> (LwF), *Riemannian Walk*<sup>27</sup> (RW) and *Modeling the Background*<sup>28</sup> (MiB). We hereby regard the datasets of each region of interest (hippocampus, prostate, or pulmonary emboli) as  $n$  tasks  $\mathcal{T}_1, \dots, \mathcal{T}_n$  and train the model of each use case sequentially with all respective tasks.

We quantify segmentation performance with the Dice coefficient and report backward transfer (BWT), which measures the degree of forgetting older tasks, and forward transfer (FWT), which assesses the ability to learn new knowledge.

Finally, we analyze the difference between using single- vs. multi-head architectures, briefly illustrate the importance of task orderings and provide a summary of our training times.

### Static results and inter-task performance

To put continual learning results into context, we first observe the performance of independent models trained solely on one dataset. These are illustrated in Figure 2. On the diagonal from the lower left to the upper right corner, we see static evaluations on in-distribution data. In this setting, all models for prostate and hippocampus achieve at least an 86% Dice. The lowest performance with merely a 7.3% Dice is for pulmonary embolism with the *FUMPE* dataset, well behind the results for the larger *CAD-PE* dataset with a 75% Dice.

The inter-task matrices also allow us to see how effectively each model performs on out-of-distribution data. These differences in performance are due to both the inherent dissimilarity between datasets in terms of acquisition and patient population

	Prostate				Hippocampus		
	UCL	I2CVB	ISBI	DecathProst	HarP	Dryad	DecathHip
<b>Static</b>	70.91 ( $\pm 6.02$ )	93.05 ( $\pm 0.29$ )	92.27 ( $\pm 0.26$ )	91.90 ( $\pm 0.36$ )	90.48 ( $\pm 1.71$ )	94.12 ( $\pm 0.05$ )	93.99 ( $\pm 0.45$ )
<b>Seq.</b>	85.16 ( $\pm 1.24$ )	21.04 ( $\pm 5.63$ )	93.09 ( $\pm 0.36$ )	<b>91.91 (<math>\pm 0.38</math>)</b>	20.20 ( $\pm 5.55$ )	57.19 ( $\pm 1.02$ )	90.92 ( $\pm 1.08$ )
<b>EWC</b>	<b>86.87 (<math>\pm 0.49</math>)</b>	58.53 ( $\pm 4.73$ )	88.43 ( $\pm 0.61$ )	87.79 ( $\pm 0.83$ )	88.01 ( $\pm 3.47$ )	86.09 ( $\pm 0.59$ )	31.93 ( $\pm 6.09$ )
<b>LwF</b>	85.30 ( $\pm 0.82$ )	22.89 ( $\pm 4.82$ )	92.37 ( $\pm 0.36$ )	91.48 ( $\pm 0.33$ )	3.90 ( $\pm 1.97$ )	46.00 ( $\pm 1.62$ )	90.85 ( $\pm 1.08$ )
<b>Reh.</b>	85.94 ( $\pm 0.76$ )	<b>90.64 (<math>\pm 0.77</math>)</b>	<b>93.39 (<math>\pm 0.28</math>)</b>	91.55 ( $\pm 0.34$ )	<b>88.17 (<math>\pm 3.63</math>)</b>	<b>92.07 (<math>\pm 0.15</math>)</b>	<b>91.16 (<math>\pm 1.17</math>)</b>
<b>MiB</b>	86.31 ( $\pm 0.62$ )	48.87 ( $\pm 6.55$ )	92.96 ( $\pm 0.39$ )	92.11 ( $\pm 0.27$ )	82.45 ( $\pm 2.94$ )	85.27 ( $\pm 0.32$ )	20.75 ( $\pm 6.99$ )
<b>RW</b>	84.08 ( $\pm 1.66$ )	26.51 ( $\pm 6.13$ )	93.18 ( $\pm 0.32$ )	92.07 ( $\pm 0.41$ )	7.33 ( $\pm 3.77$ )	34.87 ( $\pm 1.86$ )	91.07 ( $\pm 1.03$ )

**Table 1.** Continual learning performance as Dice coefficient. The first row shows the upper bound of training a model statically with all training data of the respective anatomy. We then see the performance of sequential training with and without (*Seq.*) several continual learning strategies (*EWC*, *LwF*, *Reh.*, *MiB* and *RW*). The Dice performance is reported of the final model (after training with all tasks).

and to model robustness caused by larger and more diverse training data. The assumption is that *if a model trained on  $\mathcal{T}_1$  is later trained on  $\mathcal{T}_2$ , the amount of forgetting for  $\mathcal{T}_1$  will be lower the more similar the data distribution and the higher the initial performance of the model on  $\mathcal{T}_2$ .*

For prostate segmentation (first heatmap), *I2CVB* is a clear outlier. In the case of hippocampus, the model trained on *HarP* performs worse on *DecathHip* and the other way around. While the *HarP* model achieves a 86% Dice on *Dryad*, the *Dryad* model only reaches 50% on *HarP*. This is likely due to the much larger size of *HarP* (see Table 5). The same does not hold for pulmonary embolism, where the model trained with the larger *CAD-PE* dataset badly fails on *FUMPE*.

### Continual learning methods

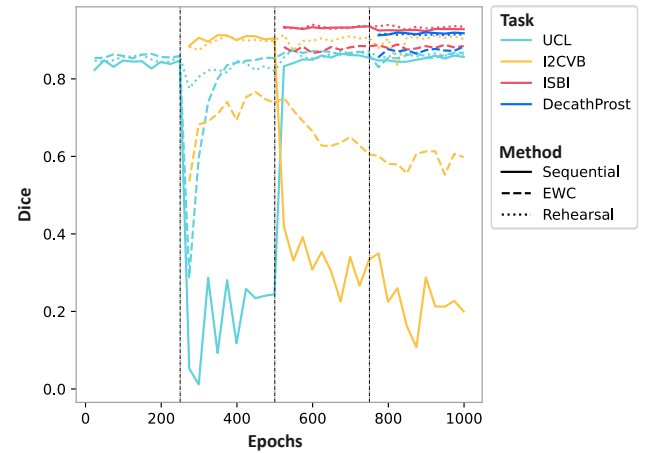
Next, we inspect the performance when models are trained in a sequential fashion, summarized in Table 1 for the prostate and hippocampus anatomies. In the first row, we report the upper bound of a static model trained with all shuffled training data from the respective anatomy. The following row shows the result of training a model sequentially in a trivial manner, and further rows are for different continual learning strategies which attempt to dampen the amount of forgetting. Reported is the Dice of the final model after training in the orders *UCL*  $\rightarrow$  *I2CVB*  $\rightarrow$  *ISBI*  $\rightarrow$  *DecathProst* and *HarP*  $\rightarrow$  *Dryad*  $\rightarrow$  *DecathHip*.

Over both anatomies, the *Rehearsal*<sup>29</sup> (*Reh.*) method is the most effective at preventing forgetting. This is consistent with previous research<sup>29</sup>. However, this strategy cannot always be used as it requires samples to be stored from previous tasks in order to interleave them in future training. This is not possible in many scenarios, where rehearsal would be an additional upper bound. In these cases, *EWC* and *RW* are good alternatives, reliably reducing the amount of forgetting. We directly illustrate the forgetting as inverse *backward transfer* in Figure 4 (y-axis), where we see that *EWC* ( $\blacktriangledown$ ) and *Rehearsal* ( $\blacktimes$ ) maintain high backward transfer scores.

Note, however, that this often comes at the cost of a loss of model plasticity, reducing the performance on later tasks. For instance, while the sequential model shows a Dice of 91.91%

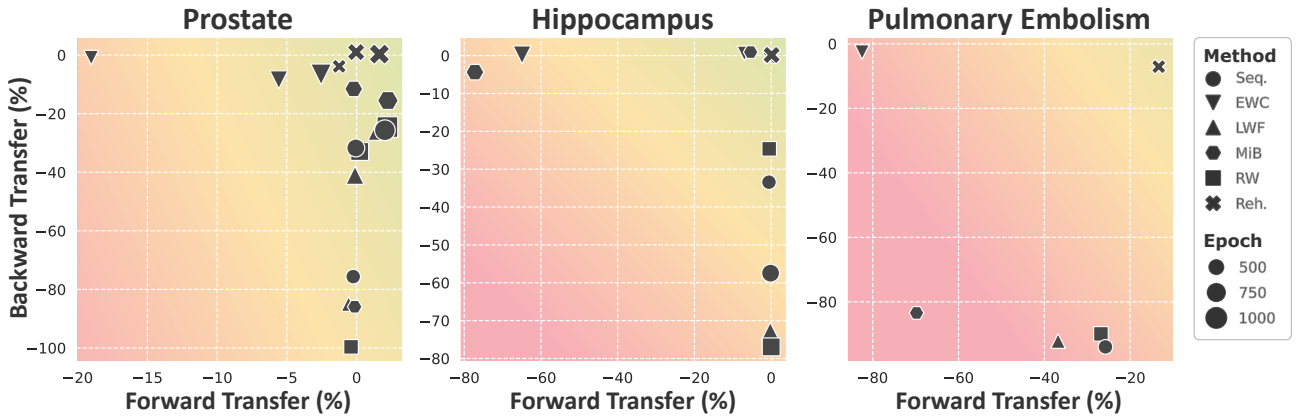
in *DecathProst* (the last task), it decreases to 87.79% for *EWC*. For hippocampus segmentation, this behavior is much more pronounced. The Dice on *DecathHip* falls from 90.92% to 31.93% for *EWC* and 20.75% for *MiB*. This is illustrated as *forward transfer* (x-axis) in Figure 4, where *EWC* shows negative values while *Rehearsal* stays close to zero.

We further analyze the behavior of trivial sequential training alongside the best-performing *Rehearsal* method and *EWC* by observing the training trajectories in Figures 3 and 5.



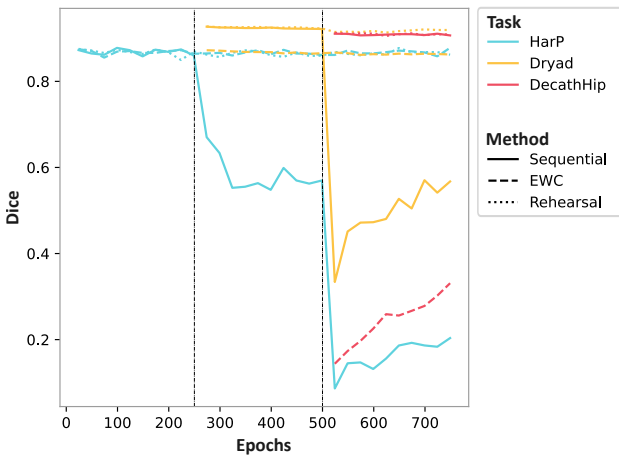
**Figure 3.** Learning trajectory for prostate segmentation. The vertical lines mark task boundaries, i.e. when training with a new task starts. Each task is displayed with a different color, and we compare trivial *Sequential* training (solid lines) to *EWC* (dashed) and *Rehearsal* (dotted).

The solid lines for sequential training mostly depict a rapid fall after task boundaries. We only see a marked recovery in Figure 3 for *UCL* (cyan) after training with *I2CVB* (second stage) is concluded. However, this is likely due to the inherent good performance of models trained with *ISBI* and *DecathProst* on *UCL* (see Figure 2). Both *Rehearsal* and *EWC* considerably reduce the amount of forgetting. However, the decreased plasticity manifesting as a negative forward transfer for *EWC* is evident, with the dashed lines of a new task often



**Figure 4.** Relative backward (y-axis) and forward (x-axis) transfer for sequential segmentation of three regions of interest, averaged over all use-case datasets. Backward transfer is the inverse forgetting and forward transfer measures how well the model adapts to future tasks. For both metrics, higher is better, and results near zero can be realistically expected.

starting below the sequential equivalents (most noticeable in Figure 5 for *DecathHip* at the third hippocampus stage).



**Figure 5.** Learning trajectory for hippocampus segmentation. A loss in model plasticity for EWC is clearly noticeable at the third training stage.

### Pulmonary embolism results and task orderings

For the segmentation of pulmonary embolism, we examine two possible orderings, i.e. *CAD-PE* followed by *FUMPE* and vice-versa. We recall that the performance on the difficult *FUMPE* dataset was only of 7.3% Dice on in-distribution test data and 7.6% on *CAD-PE* (see Figure 2). On the other hand, *CAD-PE* shows an in-distribution performance of over 75% Dice but abysmal results on *FUMPE* (1.8%).

Table 2 summarizes our findings. In the first direction (*CAD-PE*  $\rightarrow$  *FUMPE*), we see the situation where we have a model that performs well in in-distribution data. However, further training after a distribution shift causes the model to forget by 93.97% its ability to produce adequate segmentations. Such an abrupt performance decrease should be de-

tected at all costs. Fortunately, *Rehearsal* and *EWC* strategies prevent this fall in performance. If we observe the second scenario (*FUMPE*  $\rightarrow$  *CAD-PE*) we note a similar behavior. Though the absolute amount of forgetting is lower, the relative forgetting is just as significant. Unlike in the previous order *Rehearsal* does not prevent forgetting, likely due to the lower number of cases in this dataset, which means that the number of interleaved  $\mathcal{T}_1$  samples is comparatively small.

This extreme case shows how important task orderings are when comparing continual learning methods. Ideally, all orderings should be considered, but this can be computationally prohibitive when training 3-dimensional segmentation architectures. Alternatively, static in-distribution and inter-task performance results should be taken into account. We also see that *relative* backward transfer is a more intuitive metric for assessing the degree of forgetting than the absolute alternative, which requires in-tandem consideration of the base performance.

### Multi-head architectures

In previous experiments, we assumed that the entire model was sequentially trained. Continual learning is sometimes evaluated in a *multi-head* setting where the last network layer is kept task-dependent and not updated after training with its respective task<sup>15</sup>. During inference, the corresponding head is used alongside the shared body. If the task precedence is not known for a sample during inference, it can be inferred from image characteristics such as the distribution of intensity values or the ability of an autoencoder to reconstruct it<sup>19,30</sup>.

We look at how relevant this distinction is for the task of semantic segmentation by leaving the last layer task-independent. Table 3 displays the performance delta of using the “correct” head vs. that of the final model. We observe that in most cases the difference is minimal and there is no consistent pattern regarding which head is preferable. Only for *LwF* does using the correct head significantly deteriorate performance. These results indicate that, for segmentation,

	$\mathcal{T}_1 = \text{CAD-PE} \rightarrow \mathcal{T}_2 = \text{FUMPE}$				$\mathcal{T}_1 = \text{FUMPE} \rightarrow \mathcal{T}_2 = \text{CAD-PE}$			
	FUMPE		CAD-PE		CAD-PE		FUMPE	
	Dice	Dice	BWT	BWT (%)	Dice	Dice	BWT	BWT (%)
<b>Static</b>	16.81 ( $\pm 7.37$ )	68.86 ( $\pm 17.78$ )	-	-	68.86 ( $\pm 17.78$ )	16.81 ( $\pm 7.37$ )	-	-
<b>Seq.</b>	5.44 ( $\pm 3.34$ )	4.49 ( $\pm 5.31$ )	-69.99	-93.97	73.12 ( $\pm 14.72$ )	1.83 ( $\pm 2.79$ )	-8.97	-83.06
<b>EWC</b>	1.28 ( $\pm 1.51$ )	<b>72.44</b> ( $\pm 13.31$ )	<b>-1.90</b>	<b>-2.55</b>	9.90 ( $\pm 11.60$ )	<b>8.77</b> ( $\pm 4.08$ )	<b>0.50</b>	<b>6.00</b>
<b>LwF</b>	4.63 ( $\pm 3.42$ )	5.97 ( $\pm 8.16$ )	-70.17	-92.16	<b>74.34</b> ( $\pm 13.51$ )	1.86 ( $\pm 2.87$ )	-9.91	-84.20
<b>Reh.</b>	<b>6.35</b> ( $\pm 7.92$ )	69.94 ( $\pm 14.17$ )	-5.33	-7.08	72.32 ( $\pm 16.98$ )	3.81 ( $\pm 2.63$ )	-5.99	-61.11
<b>MiB</b>	2.21 ( $\pm 2.46$ )	12.52 ( $\pm 8.90$ )	-63.08	-83.44	66.95 ( $\pm 16.10$ )	0.32 ( $\pm 0.33$ )	-14.37	-97.83
<b>RW</b>	5.36 ( $\pm 4.54$ )	7.56 ( $\pm 11.52$ )	-67.65	-89.95	72.97 ( $\pm 13.99$ )	1.86 ( $\pm 2.75$ )	-4.05	-68.54

**Table 2.** Performance of sequential pulmonary embolism segmentation on two different orderings. The Dice is reported for the final model (after training with  $\mathcal{T}_2$ ). For  $\mathcal{T}_1$ , the absolute and relative backward transfer (BWT) is also reported.

	UCL		Prostate I2CVB $\Delta$		ISBI		Hippocampus				$\mathcal{T}_1 = \text{CAD-PE}$		$\mathcal{T}_1 = \text{FUMPE}$	
							HarP		Dryad		CAD-PE		FUMPE	
	Dice $\Delta$	IoU $\Delta$	Dice $\Delta$	IoU $\Delta$	Dice $\Delta$	IoU $\Delta$	Dice $\Delta$	IoU $\Delta$	Dice $\Delta$	IoU $\Delta$	Dice $\Delta$	IoU $\Delta$	Dice $\Delta$	IoU $\Delta$
<b>Seq.</b>	-0.01	-0.02	-0.04	-0.03	0.03	0.06	-0.05	-0.03	-0.04	-0.04	-0.00	-0.00	-0.01	-0.01
<b>EWC</b>	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.05	-0.05	0.01	0.01
<b>LwF</b>	-85.30	-74.37	-22.89	-12.99	-1.70	-2.89	-0.56	-0.29	1.32	1.13	0.01	0.01	-0.00	-0.00
<b>Reh.</b>	-0.04	-0.06	0.01	0.02	0.02	0.04	-0.02	-0.03	-0.03	-0.06	-0.03	-0.02	-0.02	-0.01
<b>MiB</b>	-1.38	-2.12	-1.15	-0.98	-0.11	-0.18	0.24	0.34	-0.68	-1.02	3.96	2.43	0.09	0.04
<b>RW</b>	0.00	0.01	-0.00	-0.00	0.03	0.05	-0.02	-0.01	-0.00	-0.00	0.08	0.05	-0.02	-0.01

**Table 3.** Performance difference of evaluating models with the “correct” (task-dependent) head parameters recorded after completing the corresponding training stage vs. using the head of the final model, i.e. after training with the last task. For pulmonary embolism, both orders are reported.

*maintaining task-specific heads does not have a significant effect and may actually decrease performance*, as the old heads are not tuned to the new parameters of the shared body.

### Qualitative evaluation

It is interesting to consider *which changes forgetting causes in segmentation masks*. Unlike image classification, segmentations may give a direct indication of *when* and *how* a model is failing. Figure 6 displays examples from the *UCL* and *HarP* datasets, which are the first tasks for the prostate and hippocampus use cases, respectively.

The first and second columns show the ground truth and the segmentation produced by the model right after finishing training with the corresponding task. Further columns show the prediction of the final model with different continual learning strategies. Like when trivially training the model in a sequential fashion (Seq at  $\mathcal{T}_n$ ), methods *LwF* and *RW* produce scattered segmentation masks with additional connected components. *EWC* maintains the integrity of the hippocampus segmentation, but not the prostate one. This is likely due to the increased rigidity of the hippocampus model, which in turn results in negative forward transfer (see Figure 4). *Rehearsal* generally maintains the correct shapes, though the prostate mask is larger than should be and includes one additional connected component. Finally, *MiB* successfully produces reasonable masks in both cases, though slightly lower-segments the prostate.

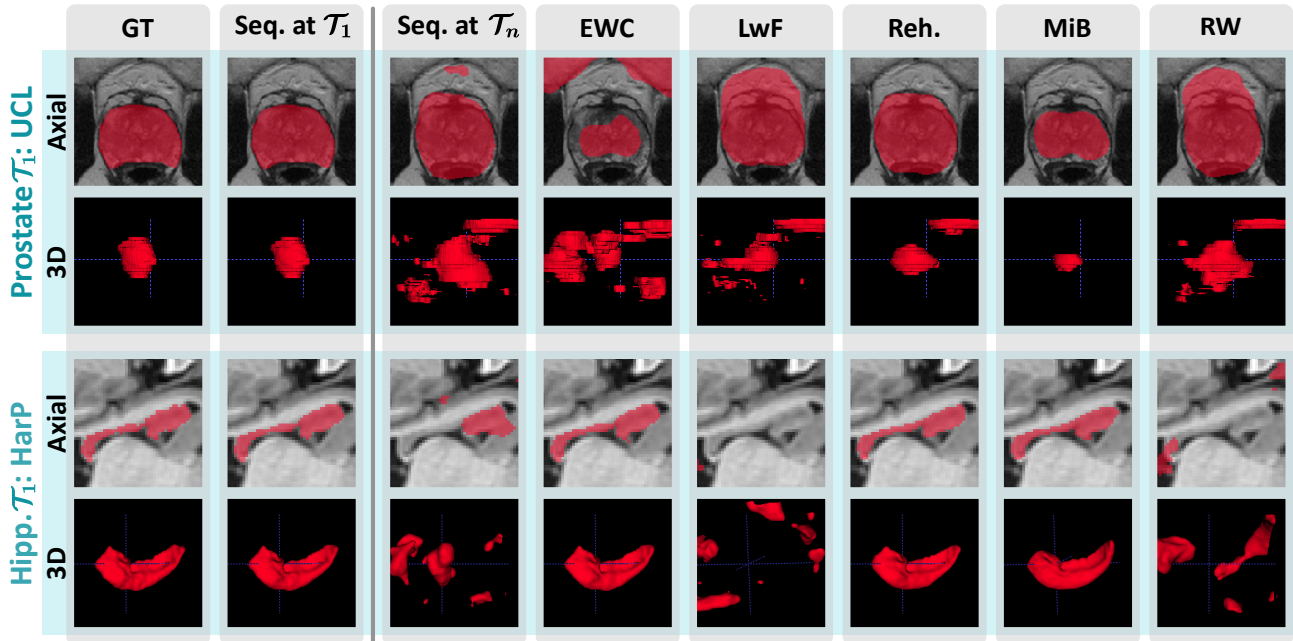
### Hardware and run times

Our experiments were carried out in a system with 8 NVIDIA Tesla T4 (16 GB) GPUs, 2 Intel Xeon Silver 4210 CPUs, and 256GB DDR4 RAM. Experiments were run in parallel, each taking up one GPU with the exception of the *LwF* experiments for the prostate and pulmonary embolism use cases. For these, 2 GPUs were used in tandem.

Table 4 provides an overview of the training times needed for one epoch for each method and anatomy. The hippocampus experiments were the fastest due to the lower resolution, while pulmonary embolism took the longest. The methods that required the longest were *LwF* and *MiB*, almost doubling the time required per epoch in several instances.

	Prostate	Hippocampus	Pulm. Emb.
<b>Seq.</b>	214.3 ( $\pm 3.1$ )	117.3 ( $\pm 20.6$ )	216.3 ( $\pm 33.3$ )
<b>EWC</b>	215.25 ( $\pm 3.9$ )	131.0 ( $\pm 5.2$ )	227.0 ( $\pm 41.6$ )
<b>LwF</b>	423.3 ( $\pm 591.5$ )	233.3 ( $\pm 252.2$ )	469.0 ( $\pm 299.6$ )
<b>Reh.</b>	206.0 ( $\pm 3.16$ )	140.3 ( $\pm 21.6$ )	223.5 ( $\pm 33.5$ )
<b>MiB</b>	365.5 ( $\pm 99.1$ )	212.0 ( $\pm 72.0$ )	318.8 ( $\pm 100.3$ )
<b>RW</b>	223.5 ( $\pm 1.3$ )	136.1 ( $\pm 1.0$ )	234.3 ( $\pm 32.1$ )

**Table 4.** Seconds required for completing one epoch of training. We report the mean and standard deviation over each anatomy, and in the case of the pulmonary embolism experiments, over both training orders.



**Figure 6.** Qualitative deterioration of segmentation performance when training models sequentially for *UCL* and *HarP*, for which we display region-of-interest crops of axial views and 3D renderings produced with *ITK-SNAP*<sup>31</sup>.

## Discussion

In dynamic clinical environments, models are needed that can adapt to changing imaging protocols and disease patterns. While the importance of continual learning for medical imaging segmentation is being recognized, our community lacks the reporting standards and benchmark datasets that researchers employ for natural image classification.

With the *Lifelong nnU-Net*, we establish a framework for the standardized evaluation of continual segmentation. We extend the popular *nnU-Net* pipeline with all components needed for training and evaluating segmentation architectures in a sequential fashion, including five popular continual learning strategies and metrics specific to continual learning paradigms.

Our evaluation across three different segmentation use cases allows us to gain valuable insights. Consistent with previous research<sup>29</sup>, *Rehearsal* leads to the best results, considerably decreasing forgetting by interleaving a subset of cases from previous tasks in the training data. Of course, a rehearsal-based strategy is only feasible if this data can be stored. For scenarios where this is not the case due to patient privacy considerations, the regularization-based *EWC* method proves to be a suitable alternative, effectively reducing forgetting though at the cost of reducing the ability of the model to adapt to new tasks. Finally, the *LwF*, *MiB* and *RW* methods do not appear to be well-suited to our setup.

One disappointing takeaway in our study is that *no method resulted in positive backward transfer (BWT)*. This is clearly illustrated in Figure 4, where we see that even the best meth-

ods only manage to prevent forgetting, reaching a BWT of zero. This means that *no knowledge acquired from later tasks improves performance on earlier tasks*. Therefore, maintaining wholly independent models and using the corresponding model during inference would outperform all explored continual learning methods. We also *only saw positive forward transfer in the prostate experiments*. This means that preceding training with earlier tasks and then fine-tuning only minimally improves performance when compared to training a model with the corresponding task from scratch.

In addition, we found that the practice of maintaining task-specific heads, common in the continual learning literature, *do not significantly affect the performance for continual segmentation in medical images*. Further studies should look into leaving a greater portion of the network task-specific.

We have identified several limitations in our study. Firstly, due to limited computational resources, we did not perform a grid search for all possible hyperparameters, such as the weight of the regularization loss for *EWC* and the memory buffer for *Rehearsal*. Instead, we used the parameters suggested in the original publications or selected with preliminary experiments using a subset of the training data and fewer iterations. We urge the reader to read the results as one possible trade-off between plasticity and rigidity, or between training time and performance. In real-world use, these hyperparameters should be tuned to obtain the desired trade-off with a separate validation set.

Secondly, we limited our study to the full-resolution patch-based 3D nnU-Net variant, which is suggested for most applications. We did not repeat our experiments on the slice-



by-slice or 3D down-sampled networks. Our evaluation also focuses on the *incremental domain learning* scenario which is most relevant in the context of medical imaging<sup>8</sup>.

Finally, as of now, there is a limited catalog of continual learning methods in the *Lifelong nnU-Net* framework. We looked to have sufficient representation of individual approaches across different strategies, and implemented a mixture of highly popular but older methods (simple *Rehearsal*, *EWC* and *LwF*) and newer approaches (*MiB* and *RW*). In the future, we hope this catalog grows both from our efforts and the contributions of other members of the community.

## Methods

An effective framework for continual image segmentation has the following requirements:

1. it has all components for achieving state-of-the-art static segmentation results and supports both two- and three-dimensional architectures (like the *nnU-Net*),
2. simplifies the evaluation of incremental domain scenarios by relying on widely accepted dataset formats and the alignment of label characteristics across datasets,
3. includes integrated evaluation logic that tracks the performance of the model for different tasks during training with appropriate metrics, and
4. supports existing state-of-the-art continual learning solutions, including the training of *multi-head models* that maintain both *shared* and *task-independent* parameters.

We start this section by introducing the three segmentation use cases that we explore, as well as our notation. We then outline how we approach each one of the requirements stated above to ensure that the *Lifelong nnU-Net* framework provides a solid foundation for medical continual learning research. Finally, we describe the continual learning methods used and briefly state details of our experimental setup.

## Datasets

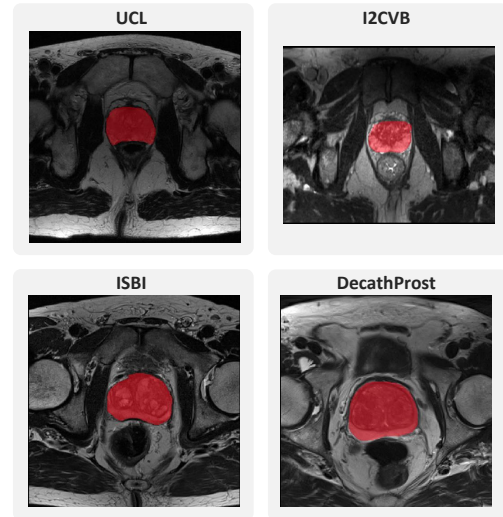
We explore the problem of continual image segmentation for three very different use cases. To ensure reproducibility, we use only openly available datasets and align the label characteristics according to the process outlined below. For each anatomy, we select an array of datasets that act as our *tasks*  $\mathcal{T}_1 \dots \mathcal{T}_n$ . Table 5 provides an overview of data and label characteristics for all datasets.

The first use case we approach is the segmentation of the prostate in T2-weighted MRIs, for which we use a corpus of four data sources. We utilize the data as provided in the *Multi-site Dataset for Prostate MRI Segmentation Challenge*<sup>32,33</sup> for sites A (*ISBI*<sup>34</sup>), C (*I2CVB*<sup>35</sup>) and D (*UCL*<sup>36</sup>). Lastly, we use the data provided as part of the *Medical Segmentation Decathlon*<sup>37</sup> (*DecathProst*). The segmentation masks contain two labels representing the peripheral zone and central gland, which we join into one *prostate* label. Prostate segmentation

Dataset	# Cases	Resolution	Spacing	ROI %	# CC
UCL	13	[24 384 384]	[3.3 0.5 0.5]	0.01	1.00
I2CVB	19	[64 384 384]	[1.3 0.5 0.4]	0.01	1.00
ISBI	30	[19 384 384]	[3.7 0.5 0.5]	0.03	1.00
DecathProst	32	[19 316 316]	[1.0 1.0 1.0]	0.03	1.00
HarP	270	[48 64 64]	[1.0 1.0 1.0]	0.01	1.60
Dryad	50	[48 64 64]	[1.0 1.0 1.0]	0.02	1.04
DecathHip	260	[36 50 35]	[1.0 1.0 1.0]	0.05	1.05
FUMPE	35	[251 512 512]	[0.9 0.6 0.6]	0.00	3.86
CAD-PE	91	[453 512 512]	[0.9 0.7 0.7]	0.00	20.46

**Table 5.** Image and label characteristics; including the number of cases, mean resolution and spacing, mean percentage of images labeled as the region-of-interest (ROI) and number of connected components (CC).

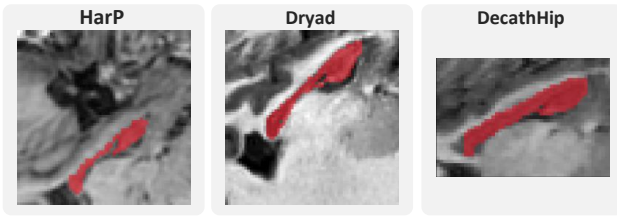
is a rather easy problem, though crucial for determining the possible location of tumorous tissue preceding a biopsy, and the shape of the prostate varies very little between different patients. Figure 7 shows examples of the four datasets.



**Figure 7.** Exemplary slices for four subjects from the prostate segmentation datasets.

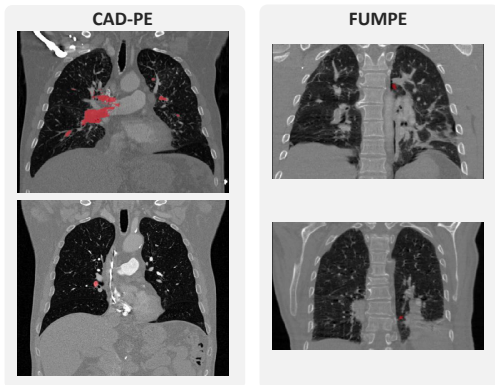
The second use case is the segmentation of the hippocampus in T1-weighted MR images, for which we include three data sources. The *Harmonized Hippocampal Protocol* data<sup>38</sup>, henceforth referred to as *HarP*, contains senior healthy subjects and patients with Alzheimer’s disease. The *Dryad*<sup>39</sup> dataset has fifty additional healthy patients. As a third data source, we use the images provided as part of the *Medical Segmentation Decathlon*<sup>37</sup> (*DecathHip*), from both healthy adults and schizophrenia patients. For the segmentation of the hippocampus, Dices of over 90% can be expected<sup>24</sup>. Exemplary image slices from all three datasets can be found in Figure 8.

Finally, we explore the segmentation of pulmonary emboli in chest CTs. This is a very complex task, as emboli can occupy few voxels and there can be multiple emboli in a scan,



**Figure 8.** Exemplary slices for three subjects from the hippocampus segmentation datasets.

possibly in different lobes. The first dataset we use is *CAD-PE*, containing 91 pulmonary angiography scans initially released for the *Computer Aided Detection for Pulmonary Embolism Challenge*<sup>40</sup>. Each embolism was originally labeled with a different class, but we merge these into one *pulmonary embolism* label. Secondly, we use the *Ferdowsi University of Mashhad’s PE (FUMPE)*<sup>41</sup> dataset, containing cases with and without embolisms and generally a lower number of emboli (see Figure 5). Exemplary slices can be observed in Figure 9.



**Figure 9.** Exemplary slices for subjects from the two pulmonary embolism segmentation datasets.

We select these three problem settings to ensure variability across modality, shape and size of the segmentation masks, and difficulty of the task at hand. Of course, our framework allows for the fast evaluation of further use cases. For all datasets, we divide 20% of the data for test purposes and maintain this split across all experiments. *We make the splits publicly available alongside our code.*

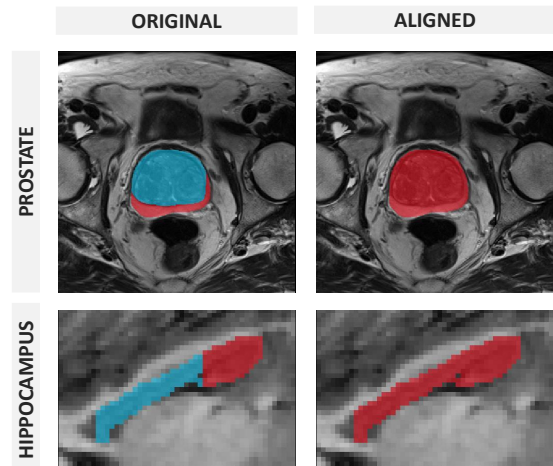
### Notation

Consider  $n$  tasks  $\mathcal{T}_1, \dots, \mathcal{T}_n$ . Model  $\mathcal{F}_2$  is trained only on the training data of task  $\mathcal{T}_2$ . Model  $\mathcal{F}_{[1,2,3]}$  was trained sequentially on tasks  $\mathcal{T}_1, \mathcal{T}_2$  and  $\mathcal{T}_3$ , in that order.  $\mathcal{F}_{\{1,2,3\}}$  is instead a *static* model, trained with shuffled training data from all three tasks. Finally, we use  $\mathcal{F}_i(\mathcal{T}_j)$  to refer to the performance of model  $\mathcal{F}_i$  applied to the test data of task  $\mathcal{T}_j$ .

### Aligning label characteristics

Very often, segmentation datasets that explore similar problems are not uniform in terms of label structure. Continual learning is only feasible if the annotations are consistent throughout datasets. Therefore, before a model can be trained in a continual fashion, a crucial pre-processing step involves *aligning label characteristics*.

Consider, for instance, the problem of prostate segmentation. Dataset  $\mathcal{T}_1$  may include annotations for the *prostate* class, distinguishing prostate voxels (which take value 1 in the segmentation mask) from the background marked with zeros. Dataset  $\mathcal{T}_2$  may instead include annotations for the central gland (label 1) and peripheral zone (label 2), two regions that together make up the prostate. Yet another dataset,  $\mathcal{T}_3$ , may include annotations for both the prostate (label 1) and bladder (label 2). We can align these labels to take up the structure of dataset  $A$  by converting annotations for labels 1 and 2 to class 1 (prostate) in dataset  $B$  and converting label 2 (bladder) to class 0 (background) for dataset  $C$ . This process is visualized in Figure 10. Of course, an alternative scenario would be *incremental label learning*, where the number of labels grows over time. In this case, one would maintain the separate *bladder* label in  $\mathcal{T}_3$ .



**Figure 10.** Alignment of label characteristics for prostate (merging the central gland and peripheral zone) and hippocampus (merging head and body).

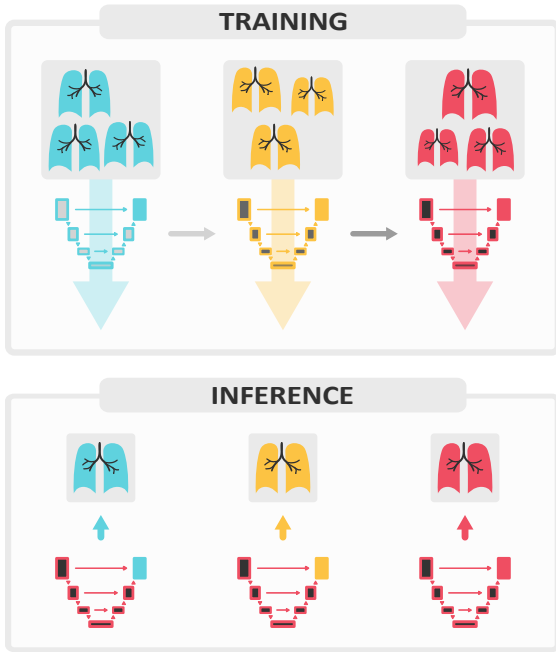
Aligning these characteristics is crucial for obtaining enough open-source data for a meaningful evaluation of different use cases. In *Lifelong nnU-Net*, we have included a pre-processing script that easily performs these steps.

### Multi-head models

The natural alternative to training a model sequentially – under our data availability constraints – is maintaining one model per task and selecting which model to use for each subject during inference. This option ensures that *no forgetting occurs*, though it leaves out any possibility for backward and forward transfer and increases the memory requirements linearly with

the number of tasks. Several continual learning methods adopt an intermediate approach: earlier layers are shared but last layers are kept task-specific<sup>25,26</sup>. The intuition is that *multi-head* models allow earlier parameters to learn from new data while the last network layers conserve task-specific information.

We implement this behavior in the *Lifelong nnU-Net* framework as visualized in Figure 11. For the first task, the training proceeds as usual. Before training takes place with the second task, the model head is replicated. Training then goes on with the shared body and the new head. This process is repeated for all tasks. During inference, a head is selected for each image and combined with the shared body. Which parameters make up the *head* is determined by the user. For the experiments on multi-head architectures, we use *seg\_outputs* as a split point.



**Figure 11.** During training, the shared body is sequentially modified while the model *head* remains task-specific. During inference, the corresponding head is merged with the final state of the shared body to extract a prediction.

### Evaluation Logic and Metrics

The nnU-Net includes methods for dataset preparation, training and performing inference. During training, the performance on a validation set is monitored. Considering the requirements of continual learning, we expanded this logic with:

- an *evaluation module* for testing on all datasets of interest, to be run after training has concluded, and
- the extended behavior of tracking the performance during training *on several different validation sets*. This gives the user insight into how the training with any task  $\mathcal{T}_i$

gradually affects the training with task  $\mathcal{T}_j$ , and allows them to export expressive training trajectories as that visualized in Figures 3 and 5.

These modifications allow for quick validation of continual learning settings and simplify the validation on out-of-distribution data without needing to store all model states.

In addition to observing the segmentation performance in the form of the Dice coefficient, we explore metrics from continual learning research that provide a more intuitive way of understanding the results. These are described in the following and visualized in Figure 12.

The primary goal of continual learning in the open world, where distribution shifts are commonplace, is to avoid overfitting to image characteristics in the last batches so that the final model can cope with samples from all seen sources. Besides avoiding the dreaded *catastrophic forgetting*, the model should ideally achieve *both backward and forward transfer*<sup>42</sup> and ensure reliable performance across all subject groups.

**Forgetting and backward transfer (BWT):** we measure the difference between the performance of a model in task  $\mathcal{T}_i$  right after training with that task and after training with further tasks. If the result is negative, this implies *forgetting* has occurred. If, instead, the result is positive, then the desirable property of *backward transfer* was achieved, e.g. training with tasks  $\mathcal{T}_{i+1}$  improves the performance on task  $\mathcal{T}_i$ .

$$BWT = \mathcal{F}_{[\dots, i, \dots]}(\mathcal{T}_i) - \mathcal{F}_{[\dots, i]}(\mathcal{T}_i) \quad (1)$$

**Forward transfer (FWT):** we calculate how advantageous the fine-tuning process is for a certain task, i.e. the difference between the continual model state right after training with task  $\mathcal{T}_i$  and model  $\mathcal{F}_i$  trained solely on task  $\mathcal{T}_i$ . A positive result implies that preceding training with data from other tasks improves the performance of the model after fine-tuning, and a negative result signifies that the model is unable to adapt to  $\mathcal{T}_i$ . This second case may occur when using certain continual learning methods that reduce model plasticity. Though other definitions consider this metric for all future tasks, we focus on the corresponding task and define:

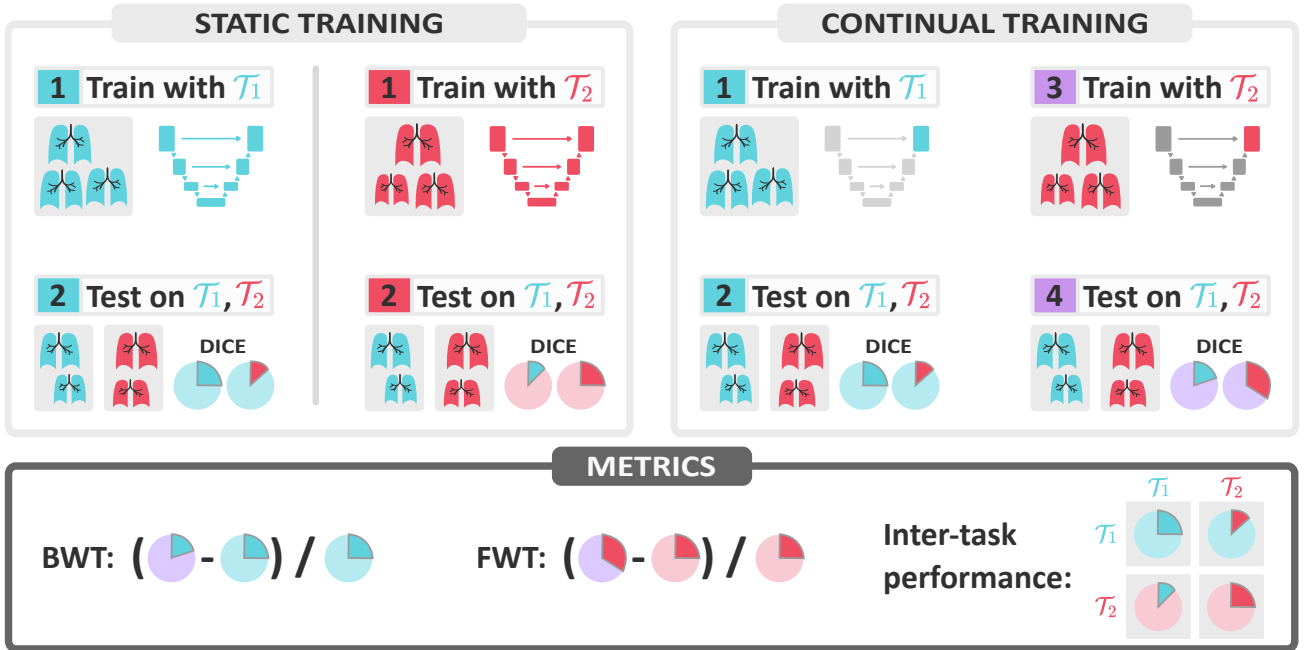
$$FWT = \mathcal{F}_{[\dots, i]}(\mathcal{T}_i) - \mathcal{F}_i(\mathcal{T}_i) \quad (2)$$

For both metrics, we report the *relative* performance change with respect to the right-hand side of the subtraction. This allows us to compare the performance across different use cases.

**Inter-task performance:** We train one separate model for each task and visualize how each model performs on the other tasks. This helps us estimate the compatibility between tasks, which should facilitate continual learning.

### Continual learning methods

**Rehearsal:** The simplest form of lifelong learning entails interleaving samples from previous tasks into the training



**Figure 12.** Continual learning evaluation metrics. We compare the performance of the segmentation model in terms of the Dice coefficient for static (left) and continual (right) training. We calculate the backward transfer (BWT) for  $\mathcal{T}_i$  as the difference in performance between the model after training with  $\mathcal{T}_i$  and after training with future tasks  $\mathcal{T}_{(j>i)}$ ; and the forward transfer (FWT) as the difference between only training the model with  $\mathcal{T}_i$  and preceding this training with previous tasks. The BWT shows the degree of forgetting and the FWT exposes the plasticity of the model or ability to learn new knowledge. Finally, the inter-task performance displays how well a model trained with each task transfers to other domains.

data. The size of the *memory buffer* determines how many of such samples are stored. The *Lifelong nnU-Net* framework allows the user to perform this type of training *with only one line of code*, specifying the tasks and size of the memory buffer. The necessary command is exemplified in Listing 1.

```

1 nnUNet_train_rehearsal 3d_fullres
2   # Specify tasks and fold
3   -t 11 12 13 -f 0
4   # Ratio of previous task data
5   -samples_in_perc 0.15
6   # Other training parameters
7   -num_epoch 250 -d 0 -save_interval 25
8   # Layer dividing body from head
9   -s seg_outputs
10  # Store evaluation results
11  --store_csv

```

**Listing 1.** Command-line directive for performing training with rehearsal. An optional *seed* argument can also be used to select samples from previous tasks in a deterministic manner.

Running other methods proceeds in a similar manner, although with different hyperparameters. Rehearsal is a very

effective strategy that consistently ensures good performance, though not admissible in settings that do not allow the storage of training samples.

**Elastic Weight Consolidation:** Regularization-based approaches assess the *importance* of each training parameter and penalize the divergence from the previous state weighted by the importance. The main difference among regularization-based methods consists in how the importance is calculated. The popular *EWC* method<sup>25</sup> utilizes the *Fisher Information Matrix*, which measures how far model outputs are from the one-hot encoded predictions. The training process is outlined in Algorithm 1.

**Learning without Forgetting:** The *LwF* method<sup>26</sup> consists of three training stages, outlined in Algorithm 2.

After the training phase for task  $\mathcal{T}_i$ , and before starting task  $\mathcal{T}_{i+1}$ , model outputs  $\mathcal{F}_{[i]}^i(\mathcal{T}_{i+1})$  are recorded and a new head is created for  $\mathcal{T}_{i+1}$ . Then, shared parameters are frozen and only the new head is trained. Finally, the shared body alongside all heads is fine-tuned. The outputs recorded in the first step are used for training previous heads.

**Riemannian Walk:** A combination of the previously introduced *EWC* with *Path Integral* forms *RW*<sup>27</sup>. The main difference to *EWC* is the online calculation of the Fisher Information Matrix. With this modification, the additional forward pass at the end of the training to obtain the Fisher values can

---

**Algorithm 1: Elastic Weight Consolidation**

---

```
Data:  $\{\mathcal{T}_i\}_{0 < i \leq nr\_tasks}$   
Args:  $\lambda_{EWC}$   
// Initialize model  $\mathcal{F}_\theta$   
1  $\theta \leftarrow \text{initialize\_model};$   
// State and importance buffers  
2  $\Theta, \Omega \leftarrow [], [];$   
// Train with first task  
3  $\theta \leftarrow \text{train}(\theta, \mathcal{T}_1);$   
4 for  $i \leftarrow 2$  to  $nr\_tasks$  do  
    // Store model states, importance  
5      $\Theta \leftarrow \Theta \cup \theta;$   
6      $\Omega \leftarrow \Omega \cup \text{Fisher}(\mathcal{T}_{i-1});$   
    // Train with EWC loss  
7      $\theta \leftarrow \text{train}(\theta, \mathcal{T}_i, \Theta, \Omega, \lambda_{EWC});$ 
```

---

---

**Algorithm 2: Learning without Forgetting**

---

```
Data:  $\{\mathcal{T}_i\}_{0 < i \leq nr\_tasks}$   
Args:  $\lambda_{LwF}$   
// Initialize body and head  
1  $\theta_B, \theta_H^1 \leftarrow \text{initialize\_model};$   
2  $\Theta_H \leftarrow \{\theta_H^1\};$   
// Train with the first task  
3  $\theta_B, \theta_H^1 \leftarrow \text{train}(\theta_B, \theta_H^1, \mathcal{T}_1);$   
4 for  $i \leftarrow 2$  to  $nr\_tasks$  do  
    // Store outputs  
5      $\mathcal{Y} \leftarrow \{\theta_H^j(\theta_B(\mathcal{T}_i))\}_{j < i};$   
    // Create and train new head  
6      $\Theta_H \leftarrow \Theta_H \cup \theta_H^i;$   
7      $\theta_H^i \leftarrow \text{train}(\theta_B, \theta_H^i, \mathcal{T}_i);$   
    // Train all parameters  
8      $\theta_B, \Theta_H \leftarrow \text{train}(\theta_B, \Theta_H, \mathcal{T}_i, \mathcal{Y}, \lambda_{LwF});$ 
```

---

be omitted.

**Modeling the Background:** The *MiB*<sup>28</sup> method – specifically developed for semantic segmentation – uses a modified cross entropy loss in combination with a knowledge distillation term. The knowledge distillation is used to force the activation of the current network  $\mathcal{F}_\theta$  to be similar to the previous network  $\mathcal{F}_{\theta_{i-1}}$ . Algorithm 3 outlines this process.

### Experimental details and hyperparameters

We train the full-resolution version of the nnU-Net which is recommended for most applications<sup>24</sup>. This is a patch-based, three dimensional network. For each of our three use cases, models are trained with every dataset for 250 epochs.

The nnU-Net automatically configures hyperparameters for the network architecture and training process - such as the number of encoding blocks, learning rate and patch size - from the training data. It is possible that these parameters differ between datasets of the same use case. In our framework,

---

**Algorithm 3: Modeling the Background**

---

```
Data:  $\{\mathcal{T}_i\}_{0 < i \leq nr\_tasks}$   
Args:  $\alpha_{MiB}, \text{lkd}_{MiB}$   
// Initialize model  $\mathcal{F}_\theta$   
1  $\theta \leftarrow \text{initialize\_model};$   
// Train with the first task  
2  $\theta \leftarrow \text{train}(\theta, \mathcal{T}_1);$   
3 for  $i \leftarrow 2$  to  $nr\_tasks$  do  
    // Extract previous model  
4      $\theta_* \leftarrow \theta;$   
    // Train with MiB loss( $\mathcal{T}_i$ ) using  
    // previous model  
5      $\theta \leftarrow \text{train}(\theta, \theta_*, \mathcal{T}_i, \alpha_{MiB}, \text{lkd}_{MiB});$ 
```

---

we always use the configuration chosen for the *first dataset*, which is the most realistic choice as in a real continual setting only this data is available when building the architecture.

For continual learning, we select hyperparameters used in previous work or which showed reasonable loss trajectories in preliminary experiments with a fraction of the epochs. For *Rehearsal*, we state the number of cases from previously seen tasks to be included in the current task to 25%. For *EWC*, we use the default value of  $\lambda = 0.4$  to weigh the importance of the regularization term. In the case of *LwF*, we set the knowledge distillation temperature to 8 for hippocampus and 64 otherwise. For *RW*,  $\lambda = 0.4$  for regularization and  $\alpha = 0.9$  for calculating the Fisher values are used. These are updated every 10th iteration. *MiB* expects two hyperparameters. The first weights the knowledge distillation loss and is set to 10 for hippocampus and 0.6 for all other cases. The second parameter is used to *hardify* the soft labels and is set to 0.9 for the hippocampus experiments and 0.75 otherwise.

We refer the reader to our code base and documentation for further details.

### Data availability

All datasets used in this work are openly available and downloading instructions can be found under the respective references.

### Code availability

Our code is available under <https://github.com/MECLabTUDA/Lifelong-nnUNet>. Upon acceptance, we will facilitate access to the trained models.

### References

1. Johnson, C. Identifying common problems in the acquisition and deployment of large-scale, safety-critical, software projects in the us and uk healthcare systems. *Saf. Sci.* **49**, 735–745 (2011).

2. Yan, W. *et al.* The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 623–631 (Springer, 2019).
3. Gonzalez, C. *et al.* Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 304–314 (Springer, 2021).
4. Liu, X. *et al.* The medical algorithmic audit. *The Lancet Digit. Heal.* (2022).
5. Food, U., Administration, D. *et al.* ‘artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd) action plan. *US Food Drug Admin., White Oak, MD, USA, Tech. Rep* **145022** (2021).
6. Rieke, N. *et al.* The future of digital health with federated learning. *NPJ digital medicine* **3**, 1–7 (2020).
7. Sheller, M. J. *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. reports* **10**, 1–12 (2020).
8. Memmel, M., Gonzalez, C. & Mukhopadhyay, A. Adversarial continual learning for multi-domain hippocampal segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, 35–45 (Springer, 2021).
9. Baweja, C., Glocker, B. & Kamnitsas, K. Towards continual learning in medical imaging. *arXiv preprint arXiv:1811.02496* (2018).
10. Perkonig, M. *et al.* Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nat. Commun.* **12**, 1–12 (2021).
11. Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z. & Mahapatra, D. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, 226–238 (Springer, 2021).
12. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Continual learning in medical devices: Fda’s action plan and beyond. *The Lancet Digit. Heal.* **3**, e337–e338 (2021).
13. Lee, C. S. & Lee, A. Y. Clinical applications of continual learning machine learning. *The Lancet Digit. Heal.* **2**, e279–e281 (2020).
14. Vokinger, K. N. & Gasser, U. Regulating ai in medicine in the united states and europe. *Nat. machine intelligence* **3**, 738–739 (2021).
15. Prabhu, A., Torr, P. H. & Dokania, P. K. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, 524–540 (Springer, 2020).
16. Mundt, M., Hong, Y. W., Pliushch, I. & Ramesh, V. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *arXiv preprint arXiv:2009.01797* (2020).
17. Hsu, Y.-C., Liu, Y.-C., Ramasamy, A. & Kira, Z. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488* (2018).
18. Lomonaco, V. *et al.* Avalanche: an end-to-end library for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3600–3610 (2021).
19. Gonzalez, C., Sakas, G. & Mukhopadhyay, A. What is wrong with continual learning in medical image segmentation? *arXiv preprint arXiv:2010.11008* (2020).
20. Michieli, U. & Zanuttigh, P. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0 (2019).
21. Cermelli, F., Mancini, M., Bulo, S. R., Ricci, E. & Caputo, B. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9233–9242 (2020).
22. Nguyen, G. *et al.* Dissecting catastrophic forgetting in continual learning by deep visualization. *arXiv preprint arXiv:2001.01578* (2020).
23. Matsumoto, A. & Yanai, K. Continual learning of image translation networks using task-dependent weight selection masks. In *ACPR (2)*, 129–142 (2019).
24. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. methods* **18**, 203–211 (2021).
25. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc. national academy sciences* **114**, 3521–3526 (2017).
26. Li, Z. & Hoiem, D. Learning without forgetting. *IEEE transactions on pattern analysis machine intelligence* **40**, 2935–2947 (2017).
27. Chaudhry, A., Dokania, P. K., Ajanthan, T. & Torr, P. H. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 532–547 (2018).
28. Cermelli, F., Mancini, M., Bulo, S. R., Ricci, E. & Caputo, B. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9233–9242 (2020).

29. Verwimp, E., De Lange, M. & Tuytelaars, T. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. *arXiv preprint arXiv:2104.07446* (2021).
30. Aljundi, R., Chakravarty, P. & Tuytelaars, T. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3366–3375 (2017).
31. Yushkevich, P. A., Gao, Y. & Gerig, G. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3342–3345 (IEEE, 2016).
32. Liu, Q., Dou, Q., Yu, L. & Heng, P. A. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging* **39**, 2713–2724 (2020).
33. Liu, Q. A Multi-site Dataset for Prostate MRI Segmentation. <https://liuquande.github.io/SAML/>.
34. N, B. *et al.* NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures, DOI: <http://doi.org/10.7937/K9/TCIA.2015.zF0vIOPv> (2015).
35. Lemaître, G. *et al.* Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput. biology medicine* **60**, 8–31 (2015).
36. Litjens, G. *et al.* Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Med. image analysis* **18**, 359–373 (2014).
37. Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR* **abs/1902.09063** (2019). [1902.09063](https://arxiv.org/abs/1902.09063).
38. Boccardi, M. *et al.* Training labels for hippocampal segmentation based on the eadc-adni harmonized hippocampal protocol. *Alzheimer's & Dementia* **11**, 175–183 (2015).
39. Kulaga-Yoskovitz, J. *et al.* Multi-contrast submillimetric 3 tesla hippocampal subfield segmentation protocol and dataset. *Sci. Data* **2**, 1–9 (2015).
40. González, G. *et al.* Computer aided detection for pulmonary embolism challenge (cad-pe). *arXiv preprint arXiv:2003.13440* (2020).
41. Masoudi, M. *et al.* A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Sci. data* **5**, 1–9 (2018).
42. Díaz-Rodríguez, N., Lomonaco, V., Filliat, D. & Maltoni, D. Don't forget, there is more than forgetting: new metrics for continual learning. In *Workshop on Continual Learning, NeurIPS 2018 (Neural Information Processing Systems)* (2018).

## Acknowledgements

This work was supported by the German Bundesministerium für Gesundheit (BMG) with grant EVA-KI [ZMVI1-2520DAT03A].

## Additional Information

The authors declare no competing interests.

---

## Discussion and limitations

Our results present interesting insights but also some disappointing facts about the current state of continual learning for medical image segmentation. Most notably, *none of the five methods we test consistently results in positive backward transfer*. Instead, methods that are successful at preventing forgetting (most notably the regularization-based *EWC*) only do so at the cost of plasticity loss.

There are also several limitations in our study. The first is that due to computational constraints we only carry out experiments for the full-resolution patch-based nnU-Net. This network achieves the best performance of the three architectures, but the nnU-Net framework also includes the possibility to train 3D downscaled and slice-by-slice networks. We did ensure when implementing the code that all methods would work with these three architectures.

Further, we study the possibility of keeping task-independent *heads*; but do not test keeping task-independent parameters for any other split points. Again, we do implement this functionality in the code, so we wish to explore further configurations in the future. Finally, we wish to establish even more benchmark datasets than the three anatomies we consider as of now, though we believe the selected tasks are sufficient for getting an overview of the benefits and downsides that different strategies for forgetting prevention signify.

I would like to highlight that this is an ongoing project. We built several other functionalities in the repository, such as extracting uncertainty scores (González et al., 2021), using domain identification oracles (González et al., 2023) and augmenting the models with transformers (Ranem et al., 2022).



---

## 12. Safe and Efficient Active Learning in Clinics

---

One new risk that continual learning introduces is the possibility that erroneous data is fed into the training set, resulting in unreliable predictions later on. Manufacturers invest considerable resources into building high-quality training sets, often employing medical professionals with high consultancy fees to produce annotations. Meanwhile, doctors working in clinical practice are overworked and do not have the time to, for instance, delineate lesions in 3D images.

In this section, we present one example of how a DL system can safely support a radiologist in detecting pulmonary emboli. We illustrate this workflow in Figure 12.1. The radiologist first performs an initial diagnosis. In parallel, a UNet model is used to segment emboli in the image. If the quality assurance system (which may include the components seen in Chapters 3 to 5) deems the prediction to be acceptable, it is supplied to the radiologist as a second opinion.

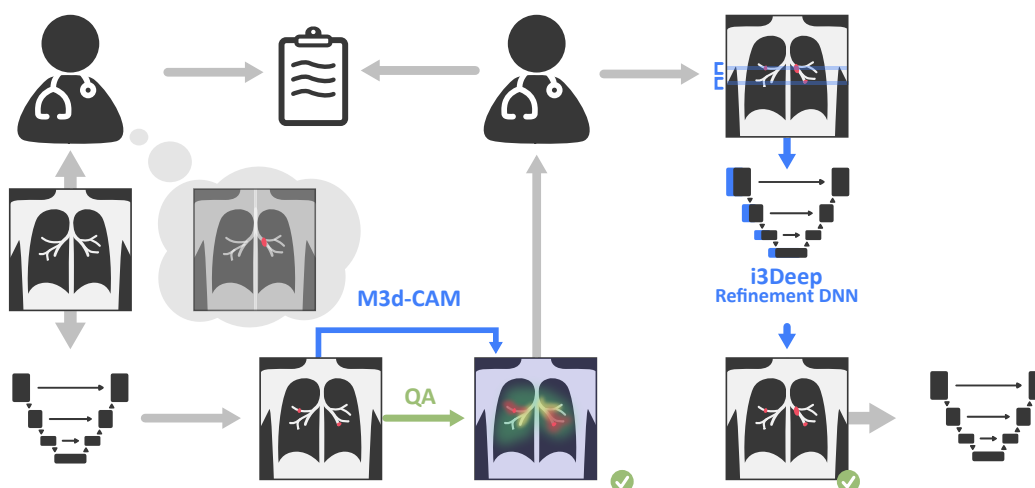


Figure 12.1.: From left to right: a radiologist receives a chest CT scan and produces an initial diagnosis. In parallel, a UNet extracts a segmentation which, if deemed confident by the QA system, is supplied to the radiologist alongside saliency maps. With the additional input, the doctor rectifies the report by adding small emboli that were initially not detected; and later corrects the segmentation mask for a portion of the volume. The image and corrected segmentation can then be used to improve model performance as an additional data sample.

A key aspect for generating trust lies in increasing transparency in the workings of the model. *Saliency maps* are the most popular DNN interpretability technique and highly sought-after by end-users. The system could supply this information together with the prediction so that the clinician can further certify the validity of the suggestion. The physician can then correct the report after looking more closely at certain areas by combining his insights with those of the model.

---

The automatically generated segmentation mask could, after being updated, be used as an additional data sample. Active learning methods, such as our *i3Deep* approach introduced in the following section, can produce high-quality annotations with minimal user input. The image with the corresponding segmentation mask can then be used for further training for an acceptable amount of work.

### A note on interpretability

In 2020, we released *M3d-CAM* (Gotkowski et al., 2021), a plug-and-play library that allows the extraction of attention/saliency maps with several methods for two- and three-dimensional data, and which works for most CNN-based *PyTorch* models (for classification or segmentation). The project is hosted at [github.com/MECLabTUDA/M3d-Cam](https://github.com/MECLabTUDA/M3d-Cam) and also integrated into *Gandlf* (Pati et al., 2021).

Our aim with *M3d-CAM* was to simplify the process of extracting saliency maps for medical imaging models. We designed the library so that it could be seamlessly integrated with existing code bases by allowing the user to *inject* the extraction mechanism in their CNN-based *PyTorch* model. The project was quite well received. At the time of writing this thesis, it has *216 stars and 26 forks*.

Nevertheless, recent work on the trustworthiness and usability of attention maps puts into question whether they are the best strategy to increase transparency of medical DNNs. Arun et al. (2021) find that they are often unreliable and difficult to interpret, though *GradCAM*, one of the methods included in *M3d-CAM*, was found to be one of the most helpful from several explored approaches. Alqaraawi et al. (2020) show that while the maps provide users a better understanding of some model features, they do not help them predict which regions of a new image the model will focus on, displaying a lack of repeatability in their behavior.

Part of this stems from the noisy nature of saliency maps, as they often highlight image regions that contribute to features that take an important role in the model but are not directly involved in making a particular decision (Kim et al., 2019). This indicates that other interoperability approaches, such as *counterfactual visual explanations* (Goyal et al., 2019) that illustrate what modifications to an image would alter the prediction, *explanatory interactive learning* where users can provide feedback to the explanations (Schramowski et al., 2020) and *model-agnostic surrogate explainers* that supply patient-specific interpretations (Kumarakulasinghe et al., 2020) may be better suited for increasing the transparency of DNN models in a way that is helpful to clinicians. On the other hand, saliency maps have helped identify cases of shortcut learning (Geirhos et al., 2020); and could potentially be utilized to prevent model deterioration over time (Patra and Noble, 2020).

What interoperability method(s) are adequate for an ML-based decision support system depends on the particularities of the problem and the intended use. An *usability analysis* should always be performed in collaboration with healthcare professionals before a product is released. Ideally, prospective studies should continuously monitor how users interact with different sources of information.

---

## 12.1. The paper: Efficient 3D interactive segmentation with i3Deep

---

One main goal of the *RACOON* (*raccoon.network*) project, which I was involved with from 2020 to early 2022, was to collect a multi-centre dataset across the network of German university hospitals that contained CT scans from Covid-19 patients and corresponding lung lesion annotations. During my involvement in the project, we developed methods to assist radiologists in producing high-quality segmentations in a reasonable time frame. The publication *i3Deep: Efficient 3D interactive segmentation with the nnU-Net* (Gotkowski et al., 2022) was one result of these efforts.

The paper was joint work with Karol Gotkowski, who thereby continued the research from his Master’s thesis (Gotkowski, 2021), Anirban Mukhopadhyay and colleagues from the interventional radiology department at the University Hospital Frankfurt. In particular, Andreas Bucher provided his counsel from a clinical perspective. Isabel Kaltenborn and Ricarda Fischbach helped prepare and annotate in-house chest CTs. The work was presented at the *Medical Imaging with Deep Learning (MIDL)* conference on July 7<sup>th</sup>, 2022, in Zürich.

### Contribution and impact

Until now we have assumed that, in continuous environments, training data becomes available over time. However, I never specified how it is actually collected. While images are naturally acquired for diagnostic purposes, in order to use them for supervised deep learning, physicians need to produce high-quality annotations in the appropriate format. This is particularly difficult to obtain for image segmentation, as manually delineating segmentation masks is highly time-consuming.

While there are existing methods for interactive image segmentation, these are often limited to lightweight architectures so they can produce real-time improvements. This is a significant downside for challenging problems where large architectures are required to obtain acceptable performance, such as lung lesion segmentation. With *i3Deep*, we propose a method that trains a *refinement network* in an offline fashion. The network accepts as input the image alongside corrected masks for only a few slices.

During the interaction with the clinician, only *two* forward passes are necessary: one for producing a pre-segmentation from the initial DNN, where slices in potential need of correction are selected using uncertainty estimation, and one through the refinement network. Correcting only one-fifth of the initial segmentation results in this manner in a 40% improvement in Dice score.

Interactive methods are highly relevant for continual learning, as they allow the generation of high-quality inputs validated by experts – which can serve as new training data – with a reasonable amount of effort. In realistic scenarios where healthcare professionals are already overworked, we must develop techniques that *assist* instead of hinder them.

# Erklärung zu Gemeinsamen Veröffentlichungen als Teil der Dissertation



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Graphisch-Interaktive Systeme (Prof. Fellner)

**Allgemeine Bestimmungen der Promotionsordnung der TU Darmstadt (8. Novelle vom 01.03.2018, §9 Absatz 5):** "Sind die zur kumulativen Dissertation vorgelegten Veröffentlichungen nicht in alleiniger Urheberschaft des Doktoranden bzw. der Doktorandin geschaffen worden, so ist eine Erklärung sowohl des Doktoranden bzw. der Doktorandin sowie aller Koautoren als auch der wissenschaftlichen Betreuerin bzw. des wissenschaftlichen Betreuers (in der Regel des bzw. der Referierenden) beizufügen, aus der sich die zu bewertenden selbständigen Leistungen anhand nachvollziehbarer Kriterien bestimmen lassen, die eine eindeutige Abgrenzung des jeweiligen Anteils ermöglichen."

Die Leistung der Doktorandin **Camila González**, betreut durch den Referenten Dr. Anirban Mukhopadhyay, bezüglich der Publikation "**i3Deep: Efficient 3D interactive segmentation with the nnU-Net**" wird folgendermaßen für ihre kumulative Dissertation festgehalten:

*The paper "i3Deep: Efficient 3D interactive segmentation with the nnU-Net" (Gotkowski et al. 2022) was published as a full research paper at the "Medical Imaging with Deep Learning (MIDL)". It constitutes a joint work of Karol Gotkowski, Camila González, Isabel Kaltenborn, Ricarda Fischbach, Andreas Bucher and Anirban Mukhopadhyay.*

*This work was supported by the Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.*

*As corresponding and leading author, K. Gotkowski led the overall research design, literature review and writing process of the paper. The choice of methodological framework and experimental setup were done by C. González and K. Gotkowski together. A. Bucher, R. Fischbach and I. Kaltenborn were responsible for collecting and processing the in-house data and reviewed the manuscript from a clinical perspective. K. Gotkowski implemented the code and conducted the experiments. K. Gotkowski and C. González contributed to the analysis of the data and results. The methodology, results and discussion were mainly written by K. Gotkowski, and C. González continuously reviewed the content and supplied feedback. The central implications of this work were mainly derived by A. Mukhopadhyay as general advisor of this work, who also contributed with continuous feedback during all phases of the paper writing process. All authors reviewed the final manuscript and agree with the use of their joint paper as part of C. González's cumulative dissertation.*

Wir sind mit der Verwendung unseres gemeinsamen Artikels als Teil der kumulativen Dissertation von **Camila González** einverstanden.

Datum: 01 / 09 / 2023      01 / 09 / 2023      01 / 12 / 2023      01 / 09 / 2023

Unterschrift:                   

Karol Gotkowski

Camila González

Isabel Kaltenborn

Ricarda Fischbach

Datum: 01 / 11 / 2023      01 / 12 / 2023

Unterschrift:       

Andreas Bucher

Anirban Mukhopadhyay

# i3Deep: Efficient 3D interactive segmentation with the nnU-Net

**Karol Gotkowski**<sup>1,2</sup>

KAROL.GOTKOWSKI@DKFZ-HEIDELBERG.DE

**Camila Gonzalez**<sup>3</sup>

CAMILA.GONZALEZ@GRIS.TU-DARMSTADT.DE

**Isabel Kaltenborn**<sup>4</sup>

ISABELJASMIN.KALTENBORN@KGU.DE

**Ricarda Fischbach**<sup>4</sup>

RICARDA.FISCHBACH@KGU.DE

**Andreas Bucher**<sup>4</sup>

ANDREASMICHAEL.BUCHER@KGU.DE

**Anirban Mukhopadhyay**<sup>3</sup>

ANIRBAN.MUKHOPADHYAY@GRIS.TU-DARMSTADT.DE

<sup>1</sup> *Applied Computer Vision Lab, Helmholtz Imaging*

<sup>2</sup> *Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg*

<sup>3</sup> *Darmstadt University of Technology, Karolinenpl. 5, 64289 Darmstadt, Germany*

<sup>4</sup> *University Hospital Frankfurt - Institut for Diagnostic and Interventional Radiology, Theodor-Stern Kai 7, 60590 Frankfurt am Main*

## Abstract

3D interactive segmentation is highly relevant in reducing the annotation time for experts. However, current methods often achieve only small segmentation improvements per interaction as lightweight models are a requirement to ensure near-realtime usage. Models with better predictive performance such as the nnU-Net cannot be employed for interactive segmentation due to their high computational demands, which result in long inference times. To solve this issue, we propose the 3D interactive segmentation framework i3Deep. Slices are selected through uncertainty estimation in an offline setting and afterwards corrected by an expert. The slices are then fed to a refinement nnU-Net, which significantly improves the global 3D segmentation from the local corrections. This approach bypasses the issue of long inference times by moving expensive computations into an offline setting that does not include the expert. For three different anatomies, our approach reduces the workload of the expert by 80.3%, while significantly improving the Dice by up to 39.5%, outperforming other state-of-the-art methods by a clear margin. Even on out-of-distribution data i3Deep is able to improve the segmentation by 19.3%.

**Keywords:** interactive segmentation, nnU-Net, uncertainty, out-of-distribution

## 1. Introduction

Manual segmentation of 3D medical data such as CT, MRI or ultrasound scans is highly time-consuming, as it often consists of hundreds of slices. Interactive segmentation reduces the workload on experts by refining the segmentation from user interactions with the goal to minimize the necessary amount and thus saving the expert time. Such methods could enable an expert to segment a CT scan with just a few clicks.

The two requirements for interactive applications are a **high predictive performance** and a **low reaction time** ( $< 1s$ ). The first enables the expert to annotate the image with much fewer interactions than when done manually, while the latter ensures the application is usable in practice. Current approaches limit the model capabilities as all their computations are performed live. To this day, no approaches exist to our knowledge that try to lift

this limitation and benefit from the much higher predictive performance of larger models. Our framework addresses this and provides an alternative by moving the expensive computations into an offline setting. Not only does this lead to fast reaction times, but also enables the use of large models, which provide much better segmentation results. Our method consists of the following steps, illustrated in Figure 1.

First, we extract both initial segmentations and uncertainties with a presegmentation nnU-Net for a subject. Based on the uncertainties, we automatically select a small number of slices with a **one-shot slice acquisition function** and send these to the expert for corrections. The corrections are then used by a **refinement nnU-Net** to improve the segmentation globally by inferring from the local corrections. Both the presegmentation and refinement nnU-Nets are trained once beforehand, with the framework solely relying on inference during the interactive segmentation process.

The expert is not involved in the presegmentation or refinement stages, which reduces the practical reaction time for the framework to zero. As a one-shot slice acquisition function is used, only a single iteration with the framework is needed to significantly improve the segmentations.

We demonstrate the effectiveness of our approach with an evaluation on the brain tumor and pancreas datasets from the Medical Segmentation Decathlon and an **out-of-distribution** in-house chest CT scan dataset with COVID-19 lesions. The code is open source and released at: <https://github.com/Karol-G/i3Deep>

## 2. Related Work

A number of interactive segmentation approaches have been proposed over the years, which we discuss in the following. An overview of the relevant methods in regards to predictive performance and reaction time is given in Table 1.

	Classical	U-Net/FCN	Konyushkova	P-Net/iW-Net	i3Deep
Predictive Performance	Low	Medium	Low	Medium	High
Reaction Time	Medium	Slow	Fast	Fast	Instantly

Table 1: Predictive performance and reaction times of interactive segmentation approaches.

**Classical methods** that are still popular today in the medical domain are Graph-Cut (Greig et al., 1989), Watershed (Meyer, 1994) and Random Walker (Grady, 2006). These methods are relatively fast even on 3D data, but have a low predictive performance by current standards.

Deep interactive segmentation approaches often outperform classical methods and most of them follow a very similar pattern of pretraining a refinement model with simulated user input and then running inference with actual expert input. However, processing higher resolution 3D images is computationally very expensive with CNNs. Therefore, approaches that employ a **U-Net** or **FCN** have slow reaction times as it is the case with Bredell et al. (2018); Li et al. (2021) and the 3D Slicer implementation of Sakinis et al. (2019).

As an alternative, other approaches use very lightweight 3D models like the **P-Net** (Wang et al., 2018, 2019b; Lei et al., 2019; Liao et al., 2020; Xu et al., 2021) or **iW-Net** (Aresta et al., 2019), which achieve a near-realtime reaction time, but have a lower predictive

performance in turn.

Besides the approaches that are task-agnostic, there are also a number of methods that are tailored to specific tasks like prostate, cell or vessel segmentation (Cheng and Liu, 2017; Koohbanani et al., 2020; Dang et al., 2022).

Other approaches used in active learning, such as by Konyushkova et al. (2015, 2019), use Boosted Trees uncertainties to find areas that should be corrected by an expert. Drawbacks of this method are the limited predictive performance and the need to retrain after every iteration.

### 3. Methodology

The i3Deep framework uses the nnU-Net (Isensee et al., 2021) for both the presegmentation and refinement model, as it has a very high predictive performance and achieves state-of-the-art results on many medical benchmarks. The training process of both models is explained in 3.1 and the inference pipeline of i3Deep is outlined in 3.2.

#### 3.1. Presegmentation & refinement nnU-Net training

We presume that a small number of subjects is already annotated, which make up the train set. Both the presegmentation and the refinement nnU-Net are trained exclusively on this train set once. The presegmentation nnU-Net is trained in a normal fashion, while the refinement nnU-Net further uses the ground truth annotations of the training set to simulate user interactions. For each image during training, slices of the ground truth are randomly chosen and all other slices are set to zero in the image volume. This modified image volume is then concatenated along the channel dimension of the image data and used as training input. When presented with corrected slices during inference, the refinement model is then able to utilize the corrections.

#### 3.2. Inference pipeline

The inference pipeline consists of a four-stage process depicted in Figure 1.

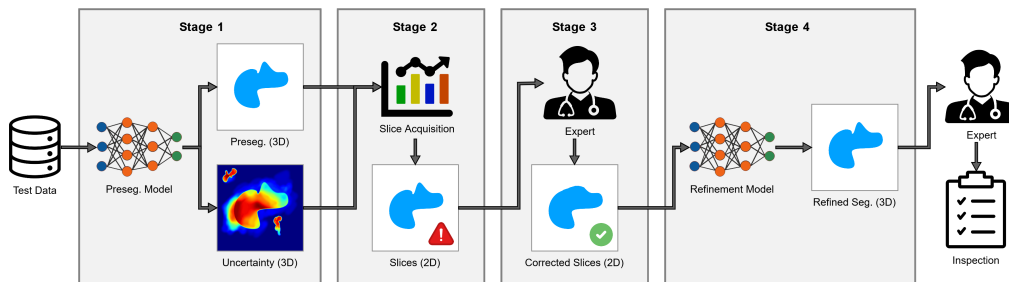


Figure 1: Overview of our proposed i3Deep framework and its four stages.

##### 3.2.1. STAGE 1: PRESEGMENTATION & UNCERTAINTY COMPUTATION

In stage one, the presegmentation model is used to run inference on new unseen subjects to provide presegmentations alongside uncertainties from the model. Estimating the uncer-

tainties for our approach can be done with multiple uncertainty predictors such as Test-Time Augmentation (Wang et al., 2019a), Monte Carlo Dropout (Gal and Ghahramani, 2016) or Deep Ensembles (Lakshminarayanan et al., 2017), which provide multiple varying predictions for an image. The voxel uncertainty inherent to the variations of these predictions is then quantified by computing their entropy. The uncertainty estimation process is expanded on in Appendix A.

### 3.2.2. STAGE 2: SLICE ACQUISITION

In stage two, a one-shot slice acquisition function selects multiple slices for each subject in axial, coronal and sagittal orientation from the 3D image based on the quantified uncertainties. The goal of this acquisition function is to select the minimum number of slices necessary to maximally improve the segmentation in a single run.

First, for each slice the sum of all uncertainty voxels is computed. Next, slices that have less uncertainty than any other slice within a minimum distance  $minDist$  are removed. This leaves only slices that are local maxima and decreases uncertainty correlation between slices. Afterwards, slices that have not enough uncertainty are removed as well, based on a  $minUncert$  parameter. Of the remaining slices, further, only a subset of  $maxSlices$  is selected that have the highest uncertainty. All three parameters are optimized after the training of the presegmentation nnU-Net once on validation data.

### 3.2.3. STAGE 3: EXPERT ANNOTATIONS

In this stage, the expert is involved in the process for the first time. The acquired slices of the previous stage are sent to the expert for correction. The expert is provided for each slice the presegmentation and subsequently corrects any mistakes they identify. We opt to let the expert choose their preferred annotation tool to enable precise corrections even on images with diffuse class borders, as it is the case with COVID-19 lesions. It is important to note that stage 1 and 2 both happen in an offline setting and the expert is only involved once these stages have been completed.

### 3.2.4. STAGE 4: REFINEMENT

In stage four, the refinement model is used to improve upon the segmentation as depicted in Figure 2. The corrected slices are projected into an empty volume back into their original

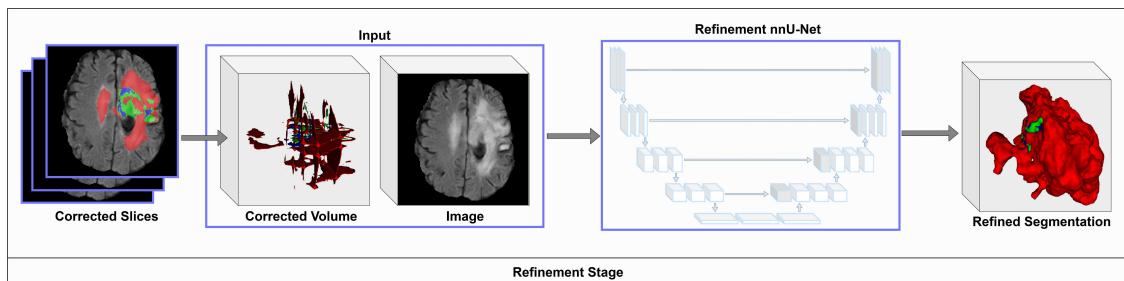


Figure 2: Inference process of the refinement nnU-Net with the corrected slices.



positions. Then this volume is concatenated with the original image and used for inference by the refinement model, which significantly improves the segmentation.

## 4. Experimental Setup

### 4.1. Datasets

We evaluate on three datasets to prove the applicability of our approach to a number of use cases. First, we use the Medical Segmentation Decathlon (MSD) **brain tumor** (Antonelli et al., 2021) dataset consisting of 484 labeled brain MRI scans with 5 MRI-modalities. The labeled classes are *edema*, *non-enhancing tumor* and *enhancing tumor* and the mean subject size of the dataset is 240x240x155 voxels. We split the dataset into a train set of 100 subjects, a validation set of 50 subjects and a test set of 334 subjects.

The MSD **pancreas** (Antonelli et al., 2021) dataset consists of 281 labeled portal-venous phase CT scans with the classes *pancreas* and *cancer* and a mean subject size of 512x512x98 voxels. Again, we split the dataset into a train set of 100 subjects, a validation set of 36 subjects and a test set of 145 subjects.

The third dataset is a **COVID-19** dataset, which consists of COVID-19 chest CT scans with the label *ground-glass opacity* (GGO). The dataset is divided into a set of subjects that are publicly available (MedSeg; Jun et al., 2020; Morozov et al., 2020) and an out-of-distribution (OOD) in-house private set to evaluate the generalizability of our approach. In total, the dataset consists of 129 subjects and a mean subject size of 1280x1280x266 voxels. The data is split into a train set of 79 subjects, a validation set of 10 subjects and an in-house OOD test set of 40 subjects.

### 4.2. Baselines

We compare our approach to other state-of-the-art 3D interactive segmentation techniques that focus on fast reaction times for the expert and can thus be used in practice. Approaches that have long reaction times such as Bredell et al. (2018); Li et al. (2021); Sakinis et al. (2019) are excluded due to their missing practicality. For the classical methods, we compare against Graph-Cut (Jirik et al., 2018; Jirik), Watershed (Skimage) and Random Walker (Skimage). For CNN-based methods, we compare against the P-Net from DeepI-GeoS (Wang et al., 2019b), which is used in most fast CNN-based approaches. We found during training that the used geodesic distance transforms from DeepI-GeoS drastically decrease the performance in our setting and thus opted to train the P-Net in the same fashion as our refinement nnU-Net instead. Further, to be able to fairly compare all baselines, they all receive the exact same corrected slices as the refinement nnU-Net from i3Deep.

### 4.3. Training details

Training of the presegmentation and refinement nnU-Nets was done in PyTorch with SGD optimizer, a learning rate of  $1e-2$ , a weight decay of  $3e-5$ , a momentum of 0.99 and 1000 epochs of training time. The P-Net used the same settings, but with grid-search optimized learning rates for the brain tumor, pancreas and COVID-19 datasets of  $1e-2$ ,  $1e-4$  and  $1e-4$ , respectively. The parameters of the acquisition function were optimized to a *minDist* of 0.0234, *maxSlices* of 12 and *minUncert* of 0.1.

## 5. Results

### 5.1. Predictive performance

We conduct an evaluation of the predictive performance in terms of Dice score performance over all datasets. The results are shown in Figure 3 and as table in Appendix B.1. Based on our uncertainty ablation study in section 5.3, we choose Deep Ensembles as the used uncertainty predictor for the presegmentation nnU-Net. However, other uncertainty methods can be used as well and are viable options for i3Deep.

Starting with the brain tumor dataset (red plots in Figure 3), we can see that the presegmentation (blue) performs acceptable for the classes *edema* and *enhancing tumor*, but rather bad for *non-enhancing tumor*. By contrast, i3Deep with nnU-Net refinement (orange) outperforms the presegmentation and all other baselines over all classes by a margin of up to 19.2%. Compared to the presegmentation, i3Deep improves the mean Dice score by 8.1%, 19.2% and 7.2%, respectively. The improvements for edema and non-enhancing tumor are lower as the Dice scores are already high for the presegmentation and thus only limited improvements are possible.

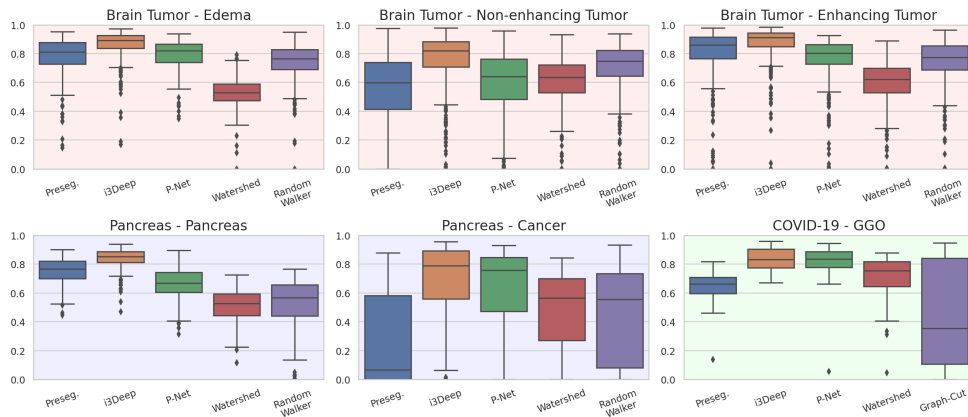


Figure 3: Box plots for different classes of the brain tumor, pancreas and COVID-19 dataset for the presegmentation, our method and all baselines.

Next, we inspect the results for the pancreas dataset (blue plots in Figure 3). For the *pancreas* class, we see again a significant improvement with i3Deep in comparison to the presegmentation by 8.4%. By contrast, all baseline methods perform significantly worse than the presegmentation, which shows their limited predictive performance. For the *cancer* class, we see that the presegmentation fails completely with a mean Dice score of only 27.4% due to the difficulty of separating the small cancer class from the pancreas class, by which it is often surrounded. In this instance, i3Deep manages to improve the segmentation by a margin of 39.5%. The P-Net improves the Dice score by 34.1%, which is also considerable. Yet, it shows again the predictive limitations of this lightweight model. The other baselines manage to improve the Dice score, but are significantly worse than i3Deep and the P-Net.

The last dataset we evaluate is the COVID-19 test set (green plot in Figure 3). It is important to note, that i3Deep has never seen any of our in-house data during training, thus

making the test data out-of-distribution (OOD) and an important benchmark for the practical usability of i3Deep. Here, the presegmentation achieves a Dice score of 64.4%, which is acceptable for OOD data. However, even though the data is OOD i3Deep still improves the Dice score by 19.3% to 83.7%, showing the applicability of our approach for real world usage. This time, the P-Net achieves a similar performance with a Dice score of 80.09%. The other baselines are again considerably worse, with Graph-Cut showing even a very high variance in terms of predictive performance.

In summary, i3Deep can improve the segmentation quality significantly in comparison to state-of-the-art baselines, while enabling the usage of models with high predictive performance such as the nnU-Net in an interactive setting.

## 5.2. Qualitative comparison

In Figure 4 a qualitative comparison of the brain tumor dataset is shown. Here, the presegmentation model fails to detect a part of the non-enhancing tumor (green) and only badly predicted the enhancing tumor (blue). By contrast, i3Deep manages to recover the missing regions almost perfectly with only minor inaccuracies for the enhancing tumor. The P-Net also recovers some of the regions, but the overall prediction lacks the same quality as that of i3Deep. The predictions for Watershed and Random Walker also recover small amounts of the missing regions, but are worse in comparison to both i3Deep and the P-Net. The pancreas and COVID-19 dataset comparison (Appendix B.2) further confirm our results.

In conclusion, all refinement models managed to recover missing lesions, yet i3Deep is the model that achieves the best segmentation in comparison to the ground truth. This shows the importance of using models with a high predictive performance in interactive settings to reliably provide segmentations of high quality for the expert.

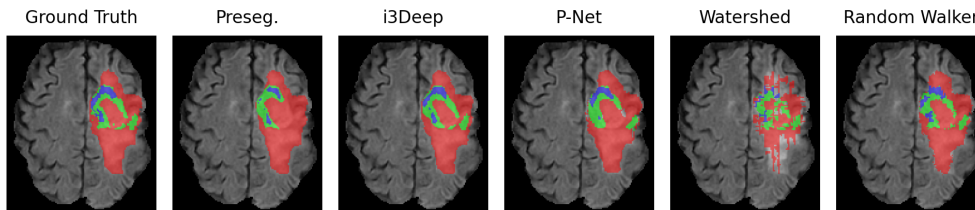


Figure 4: Qualitative comparison of the ground truth, the presegmentation, our approach and the baselines on the brain tumor dataset.

## 5.3. Uncertainty ablation

We conduct an ablation study to determine the uncertainty predictor for the presegmentation model that performs best with our approach. It is important to note that the tested uncertainty predictors are only used for the presegmentation model, as the refinement model does not need to compute uncertainties. In total, we compare the predictors Test-Time Augmentation (TTA), Monte Carlo Dropout (MC Dropout) and Deep Ensembles. The evaluation is done on all three validation datasets and measured in terms of Dice score. The results are shown in Figure 5. The Dice scores show that all predictors perform

quite similar on the brain tumor dataset with Deep Ensembles being only 0.8%, 1.3% and 0.1% better in the mean than the second best predictor on each class respectively. On the pancreas dataset the results are clearer with Deep Ensembles surpassing the second best predictor in the mean by 2.4% and 5.1%, respectively. However, Deep Ensembles perform 1.5% worse than TTA on the COVID-19 dataset. As Deep Ensembles have the best performance on most classes, we choose it as our predictor for our evaluation in section 5.1. Yet, the evaluation also shows that all three predictors are viable methods.

Further, we evaluate the predictors in terms of ECE for which the results are discussed in Appendix C.1 and reflect these results. We also evaluate the impact of using P-Net Deep Ensemble uncertainties instead of nnU-Net Deep Ensemble uncertainties in Appendix C.2. The results show that the uncertainties of both models are equally good for our approach.

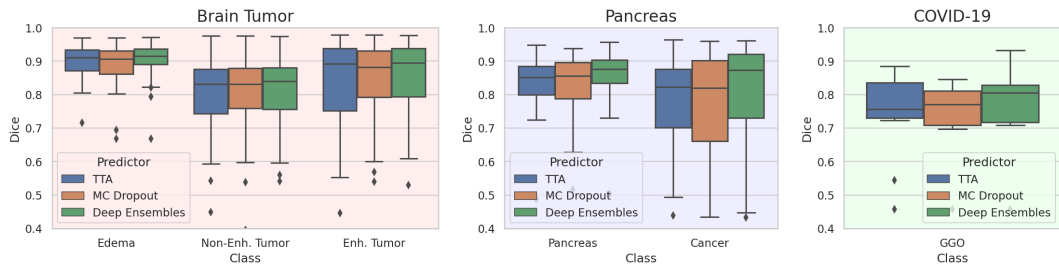


Figure 5: Comparison of the uncertainty predictors TTA, MC Dropout and Deep Ensembles on the brain tumor, pancreas and COVID-19 dataset.

#### 5.4. Annotation Ratio

To assess the expected workload reduction we propose the subject-wise *Annotation Ratio* (AR), which measures how many fewer slices need to be annotated:  $AR = \frac{|S|}{|GT_{foreground}|}$ . Here,  $|S|$  denotes the number of all selected slices and  $|GT_{foreground}|$  the number of axial ground truth slices that contain foreground annotations.

On the brain tumor dataset we achieve an AR of 20.56%, on the pancreas dataset 17.94% and on the COVID-19 dataset 20.50%. Averaged over all datasets, we achieve an AR of 19.67% meaning that an expert needs to annotate 80.33% less slices of what they would normally annotate, resulting in a significant workload reduction.

## 6. Conclusion

We introduce the interactive framework i3Deep, which enables the usage of models with a high predictive performance. i3Deep provides an expert pre-acquired slices based on uncertainties and uses the expert corrections to improve the segmentation with a refinement nnU-Net. The evaluation shows that this approach reduces the workload of the expert by 80.3%, while significantly improving the segmentations up to 39.5% and outperforming other state-of-the-art interactive methods often considerably. Even on out-of-distribution data, i3Deep is able to improve the segmentation by 19.3%. In the future, we intend to move from slices to patches and evaluate i3Deep in multiple user studies on even more anatomies and out-of-distribution datasets.

## Acknowledgments

Part of this work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

## References

- Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Goli Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, Henkjan Huisman, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbelaez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Namkug Kim, Ildoo Kim, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The Medical Segmentation Decathlon. pages 1–41, 2021. URL <http://arxiv.org/abs/2106.05735>.
- Guilherme Aresta, Colin Jacobs, Teresa Araújo, António Cunha, Isabel Ramos, Bram van Ginneken, and Aurélio Campilho. iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. volume 9, pages 1–9. Nature Publishing Group, aug 2019. doi: 10.1038/s41598-019-48004-8. URL <https://www.nature.com/articles/s41598-019-48004-8>.
- Gustav Bredell, Christine Tanner, and Ender Konukoglu. Iterative interaction training for segmentation editing networks. In Yinghuan Shi, Heung-Il Suk, and Mingxia Liu, editors, *Machine Learning in Medical Imaging*, pages 363–370, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00919-9.
- Danni Cheng and Manhua Liu. A Point Says a Lot: An Interactive Segmentation Method for MR Prostate via One-Point Labeling. volume 10541, pages 106–113, 2017. ISBN 978-3-319-67388-2. doi: 10.1007/978-3-319-67389-9. URL <http://link.springer.com/10.1007/978-3-319-67389-9>.
- Vien Ngoc Dang, Francesco Galati, Rosa Cortese, Giuseppe Di Giacomo, Viola Marconetto, Prateek Mathur, Karim Lekadir, Marco Lorenzi, Ferran Prados, and Maria A. Zuluaga. Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation. volume 75, page 102263, 2022. doi: 10.1016/j.media.2021.102263.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- Leo Grady. Random walks for image segmentation. volume 28, pages 1768–1783. IEEE, 2006.

- Dorothy M Greig, Bruce T Porteous, and Allan H Scheult. Exact maximum a posteriori estimation for binary images. volume 51, pages 271–279. Wiley Online Library, 1989.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. volume 3, pages 2130–2143, 2017. ISBN 9781510855144.
- Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. volume 18, pages 203–211. Nature Publishing Group, 2021.
- Jirik. Mjirik/imcut: 3d graph cut segmentation. pages 1–1. URL <https://github.com/mjirik/imcut>.
- Miroslav Jirik, Vladimir Lukes, Milos Zelezny, and Vaclav Liska. Multiscale graph-cut for 3d segmentation of compact objects. In *International Workshop on Combinatorial Image Analysis*, pages 227–236. Springer, 2018.
- Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. COVID-19 CT Lung and Infection Segmentation Dataset. pages 1–1. Zenodo, April 2020. doi: 10.5281/zenodo.3757476. URL <https://doi.org/10.5281/zenodo.3757476>.
- Soo Min Kang and Richard P Wildes. The n-distribution Bhattacharyya Coefficient. page 22, 2015.
- Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Introducing geometry in active learning for image segmentation. volume 2015 Inter, pages 2974–2982, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.340.
- Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Geometry in active learning for binary and multi-class image segmentation. volume 182, pages 1–16, 2019. doi: 10.1016/j.cviu.2019.01.007.
- Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: a deep learning framework for interactive segmentation of microscopic images. volume 65, page 101771. Elsevier, 2020.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. volume 2017-Decem, pages 6403–6414, 2017.
- Wenhui Lei, Huan Wang, Ran Gu, Shichuan Zhang, Shaoting Zhang, and Guotai Wang. Deepigeos-v2: deep interactive segmentation of multiple organs from head and neck images with lightweight cnns. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 61–69. Springer, 2019.

- Xiaokang Li, Mengyun Qiao, Yi Guo, Jin Zhou, Shichong Zhou, Cai Chang, and Yuanyuan Wang. Wdtiseg: One-stage interactive segmentation for breast ultrasound image using weighted distance transform and shape-aware compound loss. volume 11, page 6279. Multidisciplinary Digital Publishing Institute, jul 2021. doi: 10.3390/app11146279.
- Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9391–9399, 2020. doi: 10.1109/CVPR42600.2020.00941.
- MedSeg. Mosmeddata: Chest ct scans with covid-19 related findings dataset. pages 1–1. URL <http://medicalsegmentation.com/covid19/>.
- Fernand Meyer. Topographic distance and watershed lines. volume 38, pages 113–125. Elsevier, 1994.
- SP Morozov, AE Andreychenko, NA Pavlov, AV Vladzomyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. pages 1–4, 2020.
- Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian Binning. volume 4, pages 2901–2907, 2015. ISBN 9781577357025.
- Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring Calibration in Deep Learning. pages 0–3, 2019. URL <http://arxiv.org/abs/1904.01685>.
- Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J. Erickson. Interactive segmentation of medical images through fully convolutional neural networks. Number v, pages 1–10, 2019.
- Skimage. Module: Segmentation. pages 1–1. URL <https://scikit-image.org/docs/dev/api/skimage.segmentation.html#skimage.segmentation>.
- Guotai Wang, Wenqi Li, Maria A. Zuluaga, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. volume 37, pages 1562–1573, 2018. doi: 10.1109/TMI.2018.2791721.
- Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. volume 338, pages 34–45. Elsevier, 2019a.
- Guotai Wang, Maria A. Zuluaga, Wenqi Li, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren.

DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. volume 41, pages 1559–1572, 2019b. doi: 10.1109/TPAMI.2018.2840695.

Qisen Xu, Qian Wu, Yiqiu Hu, Bo Jin, Bin Hu, Fengping Zhu, Yuxin Li, and Xiangfeng Wang. Interactive Medical Image Segmentation with Self-Adaptive Confidence Calibration. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2021-July, pages 1–28, 2021. ISBN 9780738133669. doi: 10.1109/IJCNN52387.2021.9534435.

## Appendix A. Uncertainty estimation

Uncertainty can be estimated by multiple means and the estimation consists of two steps. First, multiple predictions need to be inferred stochastically with methods such as Test-Time Augmentation (Wang et al., 2019a), Monte Carlo Dropout (Gal and Ghahramani, 2016) or Deep Ensembles (Lakshminarayanan et al., 2017), which we refer to as *uncertainty predictors*. Second, the uncertainty from the predictions must be quantified with methods such as the entropy, variance or the bhattacharyya coefficient (Kang and Wildes, 2015), which we refer to as *uncertainty quantification*. We determine the best predictor in section 5.3, but choose entropy for the quantification as it is the most popular one and the influence of the quantification method is limited.

In this context, the *entropy* is defined as the entropy of each voxel belonging to a certain class and is based on the average of multiple predictions. Further, the entropy is divided by its information length to be within the interval of  $[0,1]$ .

For an image  $x$  with  $C$  classes and a total of  $T$  different predictions  $p_{t,c}(x)$  for each class, the entropy is defined as:

$$\overline{p_{T,c}(x)} = \frac{1}{T} \sum_{t=1}^T p_{t,c}(x) \quad (1)$$

$$H(p_{T,C}(x)) = \frac{-\sum_{c=1}^C \overline{p_{T,c}(x)} * \log(\overline{p_{T,c}(x)})}{\log(C)} \quad (2)$$

## Appendix B. Results

### B.1. Predictive performance

In this section we report the mean and standard deviation for our results of the brain tumor dataset in Table 2, the pancreas dataset in Table 3 and the COVID-19 dataset in Table 4. Dice scores marked with \* denote a  $p$ -value  $< 0.05$  when compared with the second place method. The results are the same as the one depicted in Figure 3.



Brain Tumor					
	Preseg.	i3Deep	P-Net	Watershed	Random Walker
Edema	0.784±0.128	<b>0.865±0.103*</b>	0.792±0.101	0.53±0.102	0.75±0.124
Non-E. T.	0.566±0.233	<b>0.758±0.192*</b>	0.596±0.218	0.603±0.174	0.7±0.182
Enh. T.	0.792±0.201	<b>0.864±0.158*</b>	0.751±0.186	0.598±0.155	0.74±0.175

Table 2: Mean and standard deviation Dice scores for the edema, non-enhancing tumor and enhancing tumor class of the brain tumor dataset for the presegmentation, our method and all baselines.

Pancreas					
	Preseg.	i3Deep	P-Net	Watershed	Random Walker
Pancreas	0.749±0.096	<b>0.834±0.08*</b>	0.66±0.114	0.509±0.116	0.525±0.181
Cancer	0.274±0.309	<b>0.669±0.298*</b>	0.615±0.308	0.478±0.274	0.467±0.312

Table 3: Mean and standard deviation Dice scores for the pancreas and cancer class of the pancreas dataset for the presegmentation, our method and all baselines.

COVID-19					
	Preseg.	i3Deep	P-Net	Watershed	Graph-Cut
GGO	0.644±0.125	<b>0.837±0.079</b>	0.809±0.136	0.702±0.172	0.464±0.357

Table 4: Mean and standard deviation Dice scores for the GGO class of the COVID-19 dataset for the presegmentation, our method and all baselines.

## B.2. Qualitative comparison

We continue the qualitative comparison of the pancreas and COVID-19 dataset in this section. Figure 6 shows a comparison for the pancreas dataset.

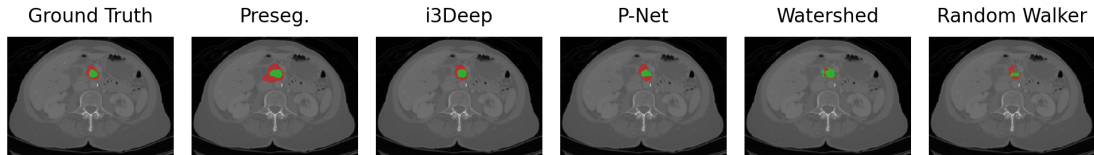


Figure 6: Qualitative comparison of the ground truth, the presegmentation, our approach and the baselines on the pancreas dataset.

Both the pancreas and the cancer class are relatively small with the pancreas class surrounding the cancer class in most subjects. It can be seen, that the presegmentation overestimated both classes. By comparison, i3Deep and the P-Net both reduced this oversegmentation,

yet i3Deep aligned the class borders overall better with the ground truth borders than the P-Net. For Watershed and Random Walker the issue of oversegmentation only increased with either the pancreas or cancer class oversegmenting the entire lesion. The comparison for the COVID-19 dataset is shown in Figure 7. Here, we see that the presegmentation missed the GGO lesions in the lower lungs, while all interactive methods were able to recover the missing lesions. However, we see again similar results with i3Deep being the most precise by not falsely segmenting the sparse small pockets of lesion free lung. The P-Net also recovered the GGO lesions for the lower lungs, but oversegmented the lung in general by segmenting the small lesion free pockets too. Again, the classical methods did not achieve the same level of refinement as i3Deep as both of them are too coarse.

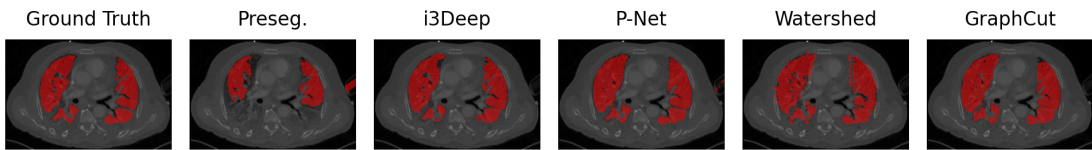


Figure 7: Qualitative comparison of the ground truth, the presegmentation, our approach and the baselines on the COVID-19 dataset.

Similar to our qualitative evaluation of the brain tumor dataset, all refinement models managed to recover missing lesion. However, i3Deep is the method that achieved the best segmentation in comparison to the ground truth, showing the importance of using models with a high predictive performance in interactive settings to reliably provide segmentations of high quality.

## Appendix C. Uncertainty ablation

### C.1. Expected Calibration Error

A common method to determine the quality of uncertainty is the *Expected Calibration Error* (ECE) (Naeini et al., 2015; Guo et al., 2017). The ECE measures the difference in expectation between confidence and accuracy to determine the miscalibration and thus the quality of the uncertainty. It divides the softmax output range of  $[0,1]$  into  $M$  multiple bins  $B_m$  of equal size and measures the accuracy and confidence of the softmax outputs that fall within each bin. A weighted average over the total number of predictions  $n$  is taken to compute a scalar miscalibration value. The ECE is formally defined as:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (3)$$

The accuracy and confidence of bin  $B_m$  are defined as follows:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \quad (4)$$

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \quad (5)$$

Here,  $\hat{y}_i$  and  $y_i$  denotes the predicted class and ground truth class for a prediction  $i$  and  $\hat{p}_i$  denotes the confidence for a prediction  $i$ .

Based on this, we evaluated the predictors TTA, MC Dropout and Deep Ensembles on the validation sets of the brain tumor, pancreas and COVID-19 dataset with the ECE measure. The results are shown in Table 5. For the brain tumor and pancreas dataset Deep Ensembles achieve the lowest calibration error and have thus the best uncertainties. By contrast, TTA performs best on the COVID-19 dataset with Deep Ensembles being slightly worse than MC Dropout. We can conclude that Deep Ensembles have probably a slight advantage over the other predictors, yet it is difficult to estimate based on the ECE how relevant that advantage is. However, in conjunction with our Dice score evaluation in section 5.3 we can conclude that this advantage is noticeable but not too significant.

Another important aspect to note is that the calibration seems to be very good based on the ECE results. However, this is most likely only partly the case as the ECE has a number of issues that are discussed in (Nixon et al., 2019) and which are especially true for 3D data, which suffers from severe class imbalance. Still, the ECE is a commonly used measure, hence us including it, but the result should always be taken with a grain of salt. For this reason, our uncertainty evaluation based on Dice score performance is more reliable.

	Brain Tumor	Pancreas	COVID-19
Deep Ensembles	0.0008	0.0004	0.0103
MC Dropout	0.0012	0.0011	0.01
TTA	0.0013	0.0006	0.0086

Table 5: The ECE results for the uncertainty predictors TTA, MC Dropout and Deep Ensembles over all three datasets.

## C.2. nnU-Net & P-Net uncertainty comparison

We evaluated the impact of using a different underlying model when using Deep Ensembles for the uncertainty computation. For this purpose, we computed uncertainties with a nnU-Net and a P-Net Deep Ensemble on the brain tumor dataset and used the resulting uncertainties to compare the predictive performance of the refinement nnU-Net and refinement P-Net. The mean Dice score result are shown in Table 6.

The performance of both nnU-Net refinement models is almost the same and independent of the uncertainty generating underlying model. The results for the refinement P-Net are similar with no significant change in model performance. However, due to the fact that the nnU-Net has a considerably better predictive performance, the refinement nnU-Net achieves a significantly better mean Dice score across all classes than the refinement P-Net.

We can conclude that there is no impact of using a different presegmentation model for uncertainty computation in our setting when using Deep Ensembles.

<b>Brain Tumor</b>				
	i3Deep (nnU-Net U.)	i3Deep (P-Net U.)	P-Net (nnU-Net U.)	P-Net (P-Net U.)
Edema	<b>0.865±0.103</b>	0.863±0.105	0.792±0.101	0.785±0.116
Non-E. T.	0.758±0.192	<b>0.770±0.183</b>	0.596±0.218	0.615±0.204
Enh. T.	<b>0.864±0.158</b>	0.863±0.1655	0.751±0.186	0.75±0.188

Table 6: Mean and standard deviation Dice scores on the brain tumor dataset for the i3Deep nnU-Net refinement model and P-Net refinement model evaluated with nnU-Net and P-Net uncertainties. The term *Uncertainties* has been denoted as  $U$ .

---

## Discussion and limitations

A central limitation of our study is that we only conduct a retrospective analysis where we simulate user interaction with ground truth annotations as “corrections”. While we had initially planned to perform a user study with our interventional radiology colleagues, this did not come to fruition due to several of the students involved in the project having other commitments and later graduating.

Additionally, the i3Deep method could not yet be validated across the network of German university hospitals due to missing key components in the data communication infrastructure. However, radiologists across Germany did receive access to our pre-trained nnU-Net model to extract pre-segmentations that they would later correct.

---

## 12.2. Conclusions and outlook

---

Despite Geoffrey Hinton’s assessment back in 2016 that deep learning would replace radiologists (Alvarado, 2022), it has become clear by now that it will not. DNNs *do* have the potential to reduce the total workload of healthcare professionals and allow for more effective utilization of hospital resources. However, this actually depends to a great extent on the involvement of human users with the system.

Figure 12.2 summarizes some of the additional tasks that working alongside DL products signify for clinicians, namely (1) curating the data, (2) monitoring the quality of the model predictions and (3) assessing whether the support systems are actually helpful. This involvement is even more crucial for continual learning algorithms, where the doctors on-site are directly involved in the data-generating process.

Advancements in quality assurance, interpretability and active learning are key for maximizing trust and minimizing the additional work that DL systems comprise. Only if sufficient medical staff is committed to using and improving the product will it translate to actual improvements in resource utilization and diagnostic accuracy.

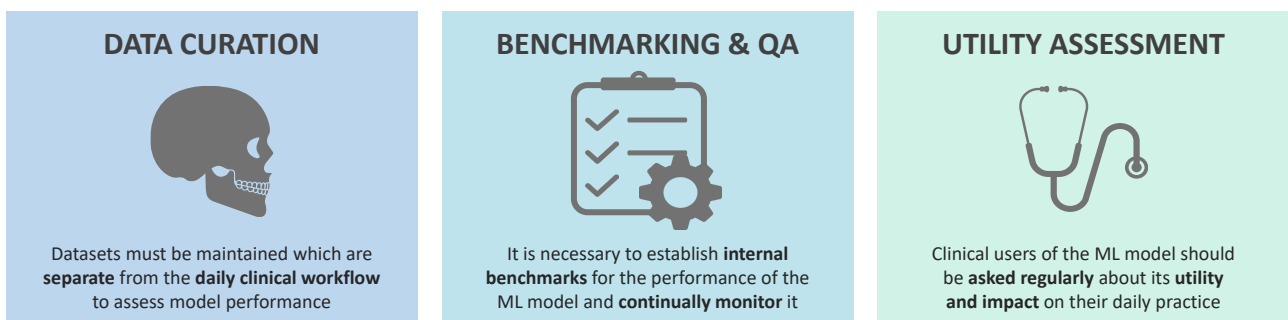


Figure 12.2.: Additional tasks that medical professionals must be involved in for ensuring the correct functioning of ML-based decision support systems.

---

## 13. Regulatory Landscape in the US and EU

---

Following technical advances in deep learning and ever-increasing healthcare costs, regulatory bodies are rushing to determine how to make safe use of the possibilities new technologies present. The landscape for ML regulation is hence rapidly evolving. In this chapter, I give an overview of directives that are currently in force and under consideration in the USA and European Union. I presented much of this content at the *EuSoMII* 2020 and *RSNA* 2021 and 2022 annual meetings, which allowed me to engage with different stakeholders, including radiologists, members of hospital administration, developers and business representatives from companies looking to market their ML software.

Software meant to aid in the diagnosis or treatment of medical conditions is deemed a *medical device* in the USA and EU. This is a different category from pharmaceutical drugs but the same group as mechanical devices such as pacemakers. If it is not part of a hardware medical device, then it is *Software as a Medical Device (SaMD)* as defined by the International Medical Device Regulators Forum (IMDRF, 2013).

There are many commonalities, but also some differences, in how SaMD regulation is approached by US authorities and EU member states. One core difference is that the USA delegates decisions to sector-specific agencies, such as the *Food and Drug Administration (FDA)* for medical devices. In contrast, manufacturers who wish to sell their products in the EU must adhere to both area-specific directives – which currently means the *Medical Device Regulation (MDR)* that came into effect on May 26<sup>th</sup>, 2021 – and cross-sectional guidelines, such as the GDPR and, in the future, an *Artificial Intelligence Act (AI Act)* that was proposed on April 21<sup>st</sup>, 2021 (Vokinger and Gasser, 2021).

Commonalities between the two models include following a *risk/benefit* model that minimizes high risks but embraces moderate risks as long as they are outweighed by potential benefits (FDA et al., 2012) and proposing a *lifecycle regulatory framework* in new directives. In both regions, said new framework would comprise (1) documentation to be submitted during the initial clearance phase outlining *what and how* modifications to the software will be implemented, (2) establishment of quality assurance mechanisms that monitor the product for its entire lifecycle, (3) data bias mitigation strategies and (4) transparency on all changes towards regulatory agencies, business partners and users.

This could enable ML models to adapt over time without obtaining clearance every time parameters are updated (as long as appropriate quality assurance mechanisms are in place). However, it would **not** permit changes that introduce new risks or change the intended use of the software.

No specific pathway has been defined and regulated in either region for continual learning systems, but official public documents suggest how these will be defined. I will start by briefly summarizing the current process for obtaining clearance in Section 13.1 and discuss new initiatives from the pertinent regulatory bodies in Section 13.2.

---

## 13.1. The road towards marketing SaMD in the USA and EU

---

In the USA there are three possible ways to receive medical device clearance from the FDA. Which process is appropriate for a device is decided by its risk categorization and whether there is a substantially similar product that is already commercially available. A comprehensive list of relevant guidelines and a list with all cleared devices are available and regularly updated on the FDA website (FDA, 2023a).

In the European Union, we find ourselves in a period of change. SaMD is now regulated under the *Medical Device Regulation (MDR)* (European Commission, 2017b), which was published in early 2017 and came into effect on May 26<sup>th</sup>, 2021, replacing the previous *Medical Devices Directive (MDD)*. A grace period for obtaining new clearance was initially stipulated for three years, so manufacturers could continue to sell products certified under the MDD until 2024. The deadline will probably be further extended due to the limited capacity of notified bodies, which are struggling to meet demand. In the following, we summarize the core relevant aspects of the MDR, the problematic associated with a lack of notified bodies, and how this calls for modern lifelong learning regulation.

### 13.1.1. Receiving approval in the USA

Software is characterized in terms of risk from *Class I* (lowest risk) to *Class III* (greatest risk). There are three main paths for receiving clearance, which may or not be pursued depending on the risks introduced by a device and whether there are sufficiently similar products that have received clearance in the past (FDA et al., 2017).

For novel devices of low to moderate risk for which no similar products have been cleared, a **De Novo** path must be followed where the manufacturer ensures the safety and effectiveness of the system (Hwang et al., 2019). After successful clearance, the FDA establishes necessary standards and controls for the new class of devices.

The road becomes simpler when there is already cleared Class I or II software with the same intended use. In this case, **510(k)** approval can be followed where the existing device serves as a reference. The new product must only show that it is *substantially equivalent* to previously-cleared devices, without the need for new clinical testing. A new piece of software can even prove similarity to other 510(k)-cleared devices, causing a chain that grows over time. This is somewhat controversial, as devices can refer to products cleared decades ago (Hwang et al., 2019), and references to devices with ongoing recalls in 510(k) submissions are associated with a significant increase in recall probability (Everhart et al., 2023).

Finally, **premarket approval (PMA)** is necessary for high-risk (Class III) medical devices which are essential for supporting life or introduce a high risk of illness or injury. This is the most stringent process and requires clinical studies and facility investigations. Makower et al. (2010) estimated the average cost of PMA clearance at 75 million US dollars, while obtaining 510(k) approval costs around 24 million. At the time of writing this thesis, there are only three ML software devices that successfully completed this pathway (FDA, 2023a).

---

### 13.1.2. The Medical Device Regulation

To enter the European market, medical devices must receive a *Conformité Européene (CE)* mark. Besides EU member states, countries that recognize this seal include *free trade association states* (Liechtenstein, Iceland, Norway, and Switzerland) and Turkey (Muehlematter et al., 2021). Switzerland historically only accepted CE-marked devices but since recently also allows FDA-certified products (Medtech, 2023).

European regulations also categorize products according to their risk, from Class I (low risk) to III (very high risk). The MDR plans to standardize this categorization, and the great majority of ML-based SaMD will fall into classes IIa and IIb (van Leeuwen et al., 2021). The *AI for Radiology* website, which collects information on available devices, currently tracks 201 products with CE marking (certified under either MDD or MDR). Of these, 66 are Class I (low risk), 117 are Class IIa, 18 are Class IIb, and none are Class III (van Leeuwen et al., 2023). From the 43 devices compliant with the MDR, 29 are Class IIa, and 14, Class IIb. Muehlematter et al. (2021) instead identify two ML software devices classified as Class III. There is yet no official website where CE-marked devices are listed, though a concrete MDR goal to increase transparency is to establish a *European database on medical devices (Eudamed)* with up-to-date information on devices in the market.

The main requirement that products should fulfill is being *safe and effective* for the expected conditions of use. The performance should be in line with the clinical state-of-the-art in accordance with the standards reached by human professionals or commercially available software. A new system for which there is already a competing product that previously received clearance should show an improvement in the benefit/risk ratio. That is to say, it should either improve performance or pose fewer risks.

Manufacturers should establish post-market *risk management* and *surveillance systems* that actively gather information during the real-world use of their product. There is an explicit wish that risk management be a continuous process that evolves throughout the product's lifecycle. The collected data should be used to regularly update the documentation, including the risk assessment sections.

There is a strong focus on the MDR in ensuring *transparency* and *traceability*. Both clinical users and patients should have access to information on the expected and real-life performance of the system and on the collected data. This is in line with GDPR principles, and it is explicitly stated that GDPR should be followed for matters concerning data protection.

There is one section of the regulation which could be potentially problematic for lifelong learning devices, namely Article 17 on “electronic programmable systems and software” (European Commission, 2017b) which states that these should ensure *repeatability, reliability, and performance*. This could imply that the system be “locked” in that the same input always produces the same output; but not necessarily. To some extent, we can consider repeatability and reliability/performance as opposing goals, as data drift causes models to fail silently. From this perspective, and considering the importance that the MDR places on a positive benefit/risk ratio and lifecycle quality monitoring, repeatability could be proven by showing that healthcare professionals assisted by the software reach the same diagnostic decisions on the initial validation set; not that every DNN in the software will produce the same output for the same input. This perspective is more in line with current ML practices, which are often not deterministic. For instance, a prediction produced by averaging several forward passes using Monte Carlo Dropout is not deterministic, yet *consistent* as long as enough passes are made.



---

## The role of notified bodies and a worrying lack of capacity

Unlike in the USA, where manufacturers are audited by a central authority, in the EU *notified bodies* take on this role <sup>1</sup>. These are third-party entities tasked with assessing conformity to present regulations. They typically start by looking over the technical documentation and, if satisfied, move on to an on-site audit where they confirm that certain quality standards are met. If certification is granted, a yearly audit should follow which can potentially result in a product being recalled if there is evidence that it is no longer compliant.

Many sections of the MDR are ambiguously defined. It is then up to the notified body, who places its seal on the audit stating that a product is safe and effective, on how these are interpreted. For instance, in the case of continual ML, if the performance of a locked system is likely to degrade over time, a notified body may determine that the risks associated with updating the model outweigh the risks of not doing so. A different one may instead insist that the system remains deterministic.

A central critique of the MDR is that there are not enough notified bodies to meet demand. All vendors need to re-certify their medical devices, and there is a multitude of new companies looking to release their products in the European market. There are currently only 36 notified bodies that can carry out this task, and only 26 more have applied to take on this role. As of October 2022, there were over 8000 applications for MDR approval, and less than 2000 had been granted (European Commission, 2023). Additionally, only auditors certified for the specific *MDR code* of a product are eligible; and notified bodies look at both the *MDR* and *EMDN* codes when deciding which products to audit (European Commission, 2017a). The *Open Regulatory* website provides an overview of notified body capacity. At the time of writing this thesis (January 2023), of 35 observed bodies, only seven were identified that are accepting new clients and offer audits in six to twelve months (Open Regulatory, 2023). It is also challenging to estimate the cost, as consultancy fees vary widely and it is difficult to estimate how long the auditing process will take (Azzouzi et al., 2022). This introduces a lot of uncertainty to companies – particularly smaller ones – looking to market their products.

Monir El Azzouzi, CEO and founder of the *Easy Medical Device* consulting firm and host of the podcast with the same name, estimates that it currently takes at least a year to bring a product to the European market from the time there is a working prototype (Azzouzi et al., 2022). For this estimate, three processes should run in parallel: (1) establishing and testing a quality management system (2) preparing all technical documentation on the working and validation of the software and (3) contacting notified bodies to secure an audit. Concerning continual or lifelong learning, he mentions that at present models are expected to be *frozen* at a certain point, as allowing adaptation makes the process of gathering the necessary documentation much more difficult. Nevertheless, he highlights that unlike for pharmaceutical drugs, the process of adhering to medical device regulations works on a case-by-case base. The most important aspect when seeking compliance is to identify the possible risks and have an appropriate risk mitigation strategy in place.

At the current rate, only around 7000 products will have been certified by the initial May 2024 deadline. There has therefore been a proposal by the European Commission to extend the grace period for MDD-certified devices until 2027 and 2028 for high and low to medium-risk devices, respectively (European Commission, 2023).

---

<sup>1</sup>Low-risk (Class I) devices do not require a notified body as manufacturers can declare conformity themselves. However, ML-based SaMD rarely falls into this category.

---

At present, any modification to cleared devices, including adapting DNNs to local environments, requires re-certification. Naturally, this puts further pressure on the few notified auditors who have experience validating ML-based software devices. A comprehensive approach to certifying lifelong learning products could alleviate the situation by substantially reducing the work of notified bodies. Currently, there is a real danger that the performance of ML software falls as the environment changes, and models can currently not be updated due to a backlog in certification applications.

---

## 13.2. Planned regulations that embrace lifelong learning

---

At present, there is no formal process for certifying ML-based software that updates models post-deployment. However, there are several initiatives that indicate how this could be approached in future regulations, which I summarize in the following sections.

### 13.2.1. The FDA's discussion paper and action plan

One of the greatest promises of AI is, according to the FDA website, “its ability to learn from real-world use and experience, and its capability to improve its performance” (FDA, 2023b). Though high-risk CL systems already function in the USA, for instance in the form of autonomous vehicles (Vokinger et al., 2021), this is not yet the case for SaMD. In May 2019, the FDA published a *discussion paper and request for feedback* showing the agency's goal to amend this situation.

The document identifies several potential benefits of continual learning, including higher accuracy, the discovery of new disease patterns, and the personalization of the system to certain patient populations or physician preferences. It also notes new risks such as the introduction of errors or biases in the data and the deterioration of performance for early data distributions (Vokinger et al., 2021).

A *locked* algorithm is defined as one that provides the same output for the same input, i.e. one that works in a deterministic fashion. In contrast, an *adaptive* algorithm may be trained post-deployment and generate a different prediction after the adaptation phase. Continuously learning algorithms are a subclass of adaptive systems that leverage real-world data post-approval. Figure 13.1 illustrates the potential of continual learning as well as new risks that it introduces into the system.

The discussion paper sketches a framework for regulatory oversight that allows model adaptation without compromising patient safety. Existing guidance documents outlying when a software modification requires a new 510(k) clearance (FDA et al., 2017) or supplemental application for devices approved through a PMA (FDA et al., 2008) are based on risk assessment. Changes must be subjected to re-approval if they introduce a new risk or modify existing risks and/or risk control mechanisms.

This includes modifications that may affect:

- the product's performance, e.g. after re-training with new data or modifying the model architecture;
- the inputs, including extending compatibility to new device models;
- the intended use, which includes changes in an expanded patient population.

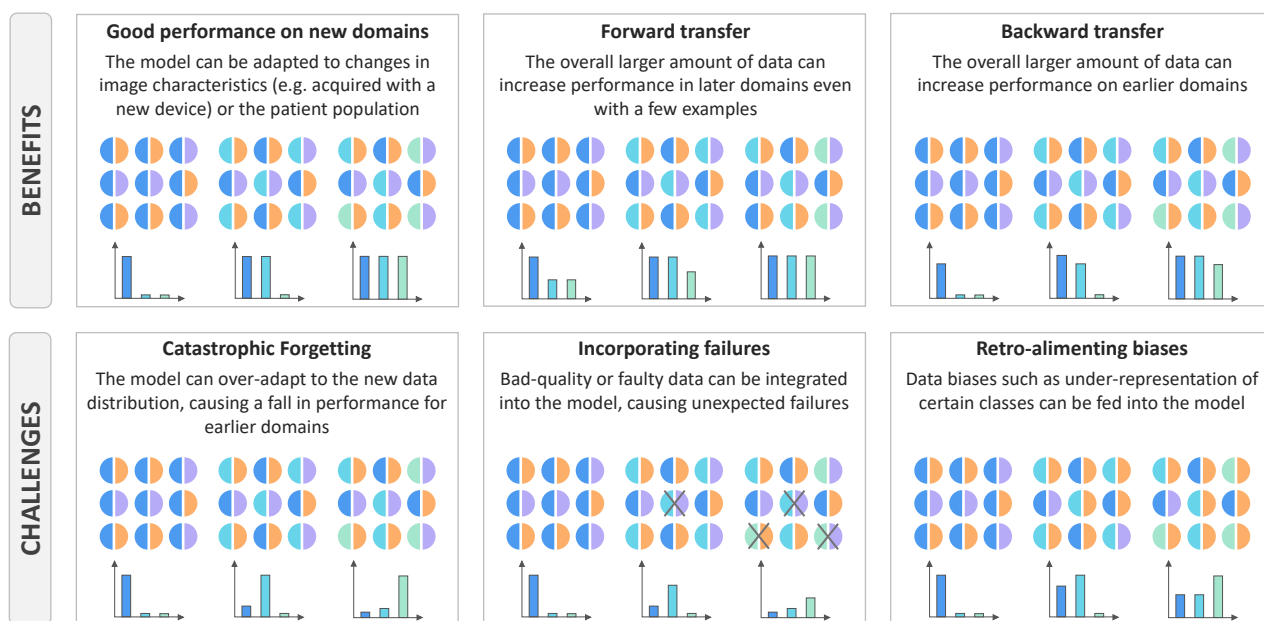


Figure 13.1.: Continuously learning ML systems adapt to new environments, obtaining higher performance on the local data distribution that generated the latest training samples. By leveraging more data overall, global performance may also increase on yet-unseen or older domains. However, it is also possible that the system forgets how to cope with earlier data or that failures and/or biases are introduced into the model.

Without diverging from the focus on risk mitigation, the new mechanism would ensure the safety and effectiveness of the product by defining from the time of initial clearance a **predetermined change control plan** stating *what* and *how* changes will occur during the product’s lifecycle. This includes *SaMD Pre-Specifications (SPS)* describing the expected changes (to inputs, performance or intended use) and an *Algorithm Change Protocol (ACP)* which specifies the data collection protocols (to ensure data is correct and unbiased), the model update strategy, and the control mechanisms to ensure the product remains safe and effective.

Future modifications that are not outlined in the approved change control plan would only require a review of the SPS and ACP. Certain changes, particularly those on the intended use that change the risk assessment of the product, would still require a traditional pre-market review under the new regulatory framework. The discussion paper describes a few examples where re-certification would and would not be appropriate.

Key elements for approving the change control plan are (a) a suitable quality monitoring strategy that considers new risks which may result from adaptation and (b) increased transparency to the FDA, collaborators of the manufacturer, clinicians, and patients through post-market performance reporting and documentation for the software version control.

The discussion paper generated a lot of interest, and feedback was gathered in the ensuing months in a public forum on the FDA website (FDA, 2023c). In February 2020, the first SaMD outlying a predetermined change control plan with anticipated modifications was cleared through the De Novo pathway. Following this development, an *action plan* was released by the FDA summarizing the stakeholder feedback and outlying clear action items to formalize guidelines in this topic (FDA et al., 2021).

---

The action plan summarizes the feedback and proposes five action items:

- Updating the proposed framework with stakeholder feedback, including, among other items, new types of modifications and a timeframe for reviewing the SPS and ACP, and issuing a first guidance on drafting predetermined change control plans;
- standardizing what entails *Good Machine Learning Practices* (GMLP), for instance related to proper validation and documentation, collaborating closely with community initiatives;
- holding a public workshop on increasing transparency and trust in ML, handling among other aspects how continuous learning can impact transparency;
- developing methodology to identify and remove algorithmic bias and evaluate robustness to changing clinical inputs; and
- clarifying expectations on Real-World Performance (RWP) monitoring. For instance, define what reference data should be used to measure performance.

In September 2022, an updated *Software Pre-certification Pilot Program* was released summarizing the findings in a pilot study that investigated good practices for monitoring SaMD real-world performance (FDA et al., 2022). One main goal of the study lay in researching whether monitoring methods could ensure safety and effectiveness over a product’s lifecycle better than the current regulatory framework. These should identify, trace and fix errors, therefore allowing for faster innovation, as is “necessary to provide reasonable assurance of safety and effectiveness of rapidly evolving devices” (FDA et al., 2022). The FDA claims being open to further feedback through the new *Digital Health Center of Excellence* (FDA et al., 2023).

### 13.2.2. The European AI Act

In the EU, products must adhere to *all* application-specific and cross-sectional regulations which are currently in force. If two directives come into conflict, the manufacturer needs to follow that which is more strictly defined or argue the selected pathway through a risk/benefit perspective. At present, products must mainly adhere to the stipulations posed by the MDR as well as other relevant directives such as GDPR. Neither of these directly addresses continual learning, though the MDR does emphasize the need for continuous monitoring.

In parallel, the *Artificial Intelligence Act (AI Act)* (European Commission, 2021) has been in the works for several years. The act does in fact describe how to treat continual learning within *high risk* applications (a category where SaMD falls). Once that directive comes into force, manufacturers could refer to that protocol to prove that planned future modifications are safe and effective.

The AI Act provides a framework for AI software covering all application areas except defense. A *regulation laying down harmonized rules on AI* was first proposed in April 2021. In a press release from December 6<sup>th</sup>, 2022, the European Council adopted the general approach of the act, indicating that further legislative action will proceed at a faster pace from now on (European Council, 2023). The act takes a risk-based approach which categorizes systems into four levels from *minimal* to *unacceptable risk*.

In several segments, the act directly addresses “AI systems that continue to learn after being placed on the market or put into service” (European Commission, 2021), which we will henceforth refer to as *CL devices*. The following declarations are particularly relevant:

- 
- Article 9 states that all high-risk AI products require a **risk management system** that is regularly updated. Over time, the system should identify and evaluate new risks and adopt measures to mitigate them. Specifically for CL devices, manufacturers should ensure post-market quality assurance mechanisms that recognize possible risks introduced by the adaptation process and react to these in a timely manner. Emphasis is also given to appropriate *record-keeping* that logs certain metrics during use, ensuring **traceability** throughout the device's lifecycle.
  - Article 15 says that high-risk AI systems must be **accurate, robust and consistent** throughout their lifecycle; and that specifically CL devices should **prevent the generation of biased outputs**. Given the direct referral to CL devices, the *consistency* requirement likely does not refer to generating the same output for a given input. Rather, the focus lies on ensuring a high level of performance over time. It is later explicitly stated that in accordance with non-discrimination laws, quality data sets should be maintained (which implies **dataset updating**) and biases should be mitigated throughout the product's lifecycle.
  - Article 43 clarifies that high-risk AI systems will need a new conformity assessment when they are *substantially modified*. It then specifically states that for CL devices, pre-planned modifications described in the technical documentation of the initial assessment **are not a substantial modification**. It is also clarified that a new conformity assessment is required when a change may affect compliance with the act or when there is a **change in the intended use**.

### 13.2.3. Commonalities and a look into a future lifecycle regulatory approach

In both regions, there is a strong focus on maximizing the benefit/risk ratio and controlling that it does not diminish over time. This implies that if the risks associated with system adaptation are outweighed by the benefits; or if not updating the system poses an additional risk, manufacturers should minimize this risk by adapting the system to the new environment.

The way this would take place is by allowing manufacturers to supply additional documentation stating which changes will occur over time and how this process will proceed, as illustrated in Figure 13.2. Changes that were specified can then occur without additional clearance. That is, of course, as long as the planned modification remains within “normal conditions of use”. That is to say that while adapting the model to work on new scanners or naturally occurring population changes is acceptable, adjusting it to widely different conditions of use (such as scans acquired with vs. without contrast agents) is not acceptable. The EU and USA also share similar objectives related to increased post-deployment quality monitoring, bias mitigation, and transparency.

Vokinger and Gasser (2021) sustain that the EU approach is more stringent, or at least defines requirements in a more specific manner. This methodology clashes with certain characteristics of the EU setting, potentially making adherence problematic. For instance, under EU regulations, medical software must maintain state-of-the-art performance throughout its lifespan across all subject groups for which the software is intended. Yet patient populations vary greatly within EU member states, and gathering geographically diverse data may be challenging. Additionally, electronic health data formats vary within states, and the GDPR poses strict stipulations for data protection. Also in terms of transparency, EU guidelines are more far-reaching, demanding that users obtain insight into the systems. Yet, for instance, there is not yet an official public website with up-to-date information on commercially available medical devices similar to the FDA's.

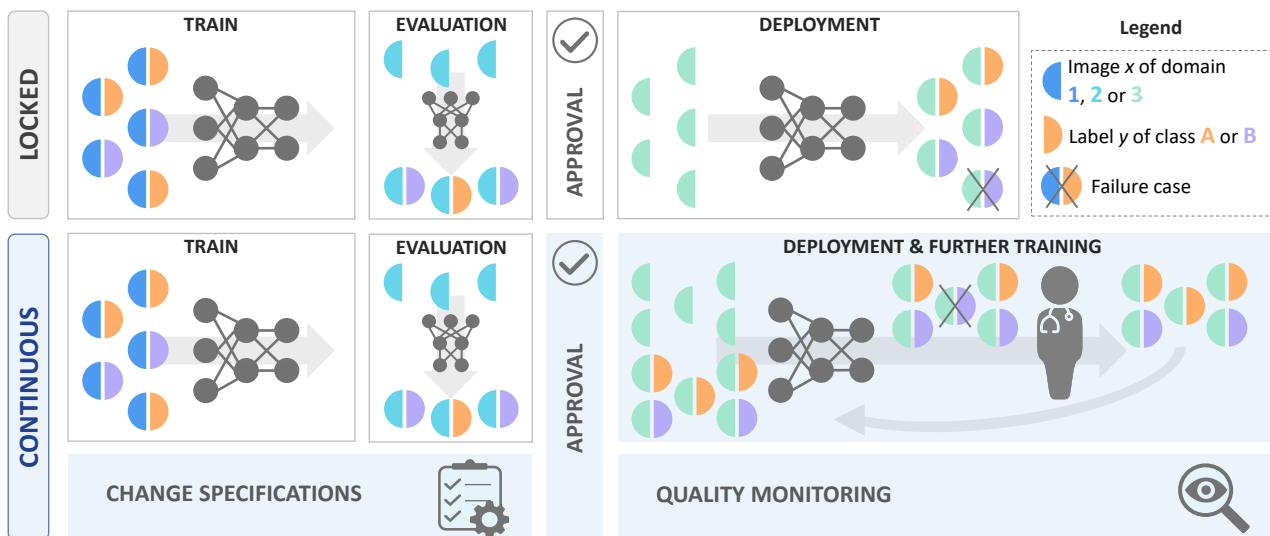


Figure 13.2.: Possible framework for regulating continual ML products. Planned modifications and pre-defined and submitted as extra documentation during initial clearance. Additional quality assurance during the product's lifecycle oversees both the outputs generated by the model and the new training data.

Muehlematter et al. (2021) instead maintain that obtaining clearance in the EU is simpler. They follow 222 approved medical devices in the USA and 240 in Europe. Of the 124 approved for both regions, 80 were first cleared according to EU standards, which could imply a simpler process to enter the European market. This is supported by the fact that 12 products that were commercially available in the EU were denied entry to the US market for not meeting safety and effectiveness standards. Manufacturers surveyed by Makower et al. (2010) also state that the European process is faster and less expensive.

Besides the cases explored in more detail of the USA and EU, other states including Canada and the UK are drafting regulations that may soon describe how best to handle continual ML products (Smith and Severn, 2022). I focused on the FDA and EU cases in part as products certified under these regulations are permitted in numerous countries (Muehlematter et al., 2021; Smith and Severn, 2022).

### 13.3. Conclusions and outlook

When I approach regulatory consultants or companies selling medical ML software on the topic of continual learning, the response I usually receive is that we are far from moving to a lifecycle regulatory protocol that allows such post-market modifications. However, this neglects how important many aspects of continual learning research are *today*.

DNNs may not be updated in an *online* manner, immediately integrating new data and annotations. But companies *do* collect data – often in the form of corrections made by healthcare professionals – with the hope of improving their models over time by adding new findings, adjusting them to new acquisition practices, or preparing them to be deployed in new geographic regions. These changes are introduced with new product versions in discrete intervals (as we illustrate in Figure 13.3), but they are still *continual*

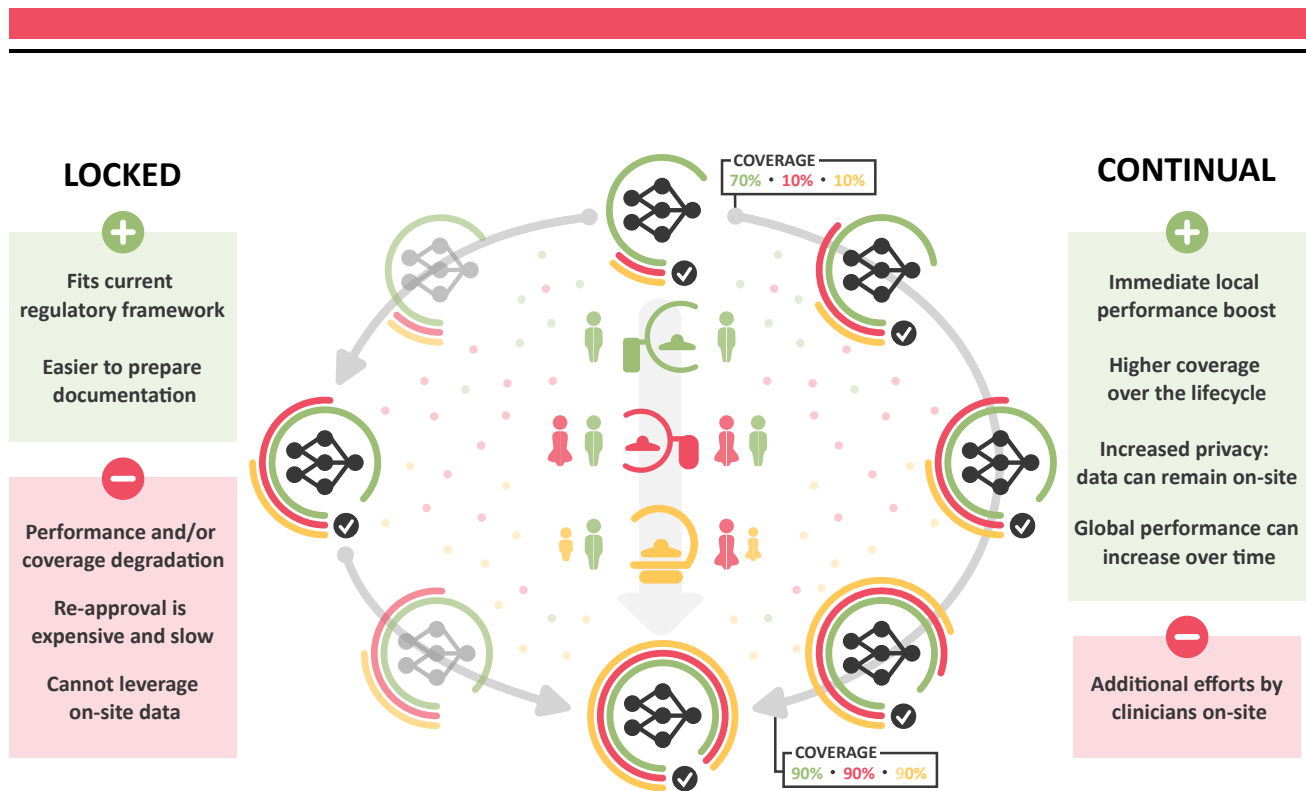


Figure 13.3.: Advantages and disadvantages of locked vs. continual learning systems with respect to resource utilization and model performance.

in the sense that the system keeps on learning post-deployment. In fact, most continual learning research divides training into stages and evaluates the performance on different data distributions after each stage. That the adaptation process is divided into segments does not make it *static*, unless there are no plans to improve the product over time.

A lot of the questions I posed, such as how can we cope with changing data availability and what are appropriate evaluation datasets and metrics, are essential right now beyond the specific subfield of continual learning research. For instance, when we have new labeled cases but lost access to some of the training data, should we re-train our model or not? If the performance of our model has decreased on the initial validation set but increased for more recent data, should the update be rolled out? What happens when the detection of several findings improves, but that of one finding deteriorates? There is a discourse around these issues that I believe is missing from public policy and medical imaging circles. Regulatory bodies often consider existing standards established by researchers and members from industry, so we can help draft these answers (Azzouzi et al., 2022).

I am fully aware of the dangers of unmonitored model updating and do not endorse this avenue. The way I understand the standing and upcoming regulatory documents for SaMD, each model update that is rolled out will need to be tracked and documented with accompanying validation results. My hope is that when the validation shows an increase in safety and/or effectiveness, a model update can be carried out *faster*, without waiting for months for regulatory authorities or notified bodies to have sufficient capacity. For this, unified standards need to be established on what signifies an increase in effectiveness. The majority of medical ML-based software has only been commercially available for a few years or less, so many companies have not yet rolled out any model updates. But all evidence on the effects of data drift on DNNs suggests that the lifespan of static models is short.

There is some misconception that current regulations are, if anything, overly conservative/risk-averse and that continual learning is avoided because it poses too big of a risk. Yet we have seen that there are

---

simple techniques – such as expansion-based multi-model approaches – that can be flexibly applied to most modern architectures and do not cause performance degradation on older data distributions. And there are valid concerns on the current process of ML software clearance. For instance, Wu et al. (2021) investigate 130 certified devices and find that 126 only went through retrospective studies, and 93 did not report any multi-center validation. From 54 high-risk devices, none showed prospective evidence. van Leeuwen et al. (2021), similarly, report that for 100 commercially available AI products, 192 from 237 studies were retrospective, and only 71 released multi-site results.

The risks of static ML systems that we explored in previous chapters – such as the gradual deterioration of performance caused by data drift – cannot be overlooked. Changes in acquisition practices and reconstruction algorithms are also accelerating, which means that the data drift problem will manifest faster than expected. Following the benefit/risk perspective taken by both the FDA and European Commission, and considering the fact that DNNs fail silently and many commercially available software systems lack sufficient prospective validation, further postponing a lifecycle regulatory approach that allows controlled model updates and requires sufficient QA poses too high a risk. If we act fast enough, a regulatory framework that is more beneficial to healthcare professionals and patients may be in place before the next medical condition spreads, pushing healthcare resources to a limit. We could then leverage DL advances for more effective action across the globe.





## **Part IV.**

# **Summary and Future Perspectives**



---

Whenever a colleague working in a radiology department mentions that a deep learning solution is being actively used, my first reaction is to be pleasantly surprised. Most DL-based products only recently came into the market, but there are now several hundred devices cleared for clinical use in the EU and USA (van Leeuwen et al., 2021). Unfortunately, the initial excitement drops when I hear about the software committing blatant mistakes, and I cannot help but think about the many failures that are going undetected. How can this happen when we expect regulatory authorities to only approve systems that are proven to be *safe and effective*?

I believe the core of the problem lies in the fact that most DL evaluations, both in research publications and for commercial products, are made in static, close-world environments where most factors behind the data distribution are accounted for. In fact, van Leeuwen et al. (2021) find that 192 from 237 studies performed by manufacturers for obtaining certification were retrospective. Wu et al. (2021) state that from 130 observed certified devices, *only four* overall went through prospective studies, and this includes the 54 high-risk devices. Regulatory bodies have identified this problem and are moving towards a *lifecycle approach* where manufacturers and clinical users must continuously monitor devices.

In this thesis, I take a holistic view at the problem of designing systems for the clinical open world so that they remain *safe and effective over time*. Substantial performance degradations of DNNs are in no way surprising. Even if manufacturers train their systems with large, heterogeneous datasets, with time *data drift* will inevitably take place and the system will begin to fail. In Chapter 1, I explore what changes cause a shift in the data distribution. These can be related to both the image acquisition practices and the patient population. In fact, there are often several factors that introduce changes in parallel, even within the same institution.

This performance deterioration is particularly problematic because DNN failures tend to be *silent*. The first step is, therefore, to detect when a prediction is uncertain. In Chapter 3, I summarize OOD detection methods I developed for this purpose that either monitor the value of a *proxy loss* (González and Mukhopadhyay, 2021) or calculate the distance to training features in a low-dimensional latent space (González et al., 2021). The second strategy works reliably across a number of use cases (González et al., 2022a) and is applicable to virtually any DNN model without needing re-training. This is a significant advantage, as models are mostly chosen for their target-task performance and not for secondary goals such as outlier detection. In future work, I wish to research ways to better capture properties of the latent space and identify *in what ways* a sample is different from the training data.

Uncertainty estimation can also help us identify low-confident predictions. In Chapter 4, I describe a method to increase the diversity between members in a deep ensemble (Mehrtens et al., 2022). I also show how domain knowledge can be used to find failure cases (González et al., 2022b) in Chapter 5. I do not believe that any one method should be completely trusted. Instead, a quality monitoring system as that suggested for a lifecycle regulatory approach (FDA et al., 2019) should quantify multiple metrics in tandem.

Effectively detecting failures can allow the model to be utilized for a portion of the cases. However, our objective is to *maintain good performance for the majority of the data*. Domain adaptation and OOD generalization methods pursue this goal with some success. However, as I mention in Chapter 6, robustness is only increased up to a certain level and at the cost of considerable computational overhead.

We know by now that a model is only as good as the data it was trained on. For a DNN to maintain stable performance through its lifecycle, it inevitably needs to be trained with new samples. *Continual learning* looks to adapt models to changes in the environment while preserving previous knowledge. As I explain in Chapter 7, most methods attempt to find an appropriate trade-off between *rigidity* and *plasticity*.

---

There is, however, a simple solution that maximizes both these objectives, namely *expansion-based approaches* that maintain several parametrizations of the model and employ the most appropriate one during inference. I present two methods in Chapter 8 that apply this concept to *task-agnostic* scenarios where no information is provided about the data-generating process. The first introduces an *oracle* that trains one autoencoder network per domain and, during testing, selects the model corresponding to the domain with the smallest error (González et al., 2023). The second method, *ODEx*, leverages OOD detection and works even in situations where there are no rigid domain boundaries during training (González et al., 2022d). A central challenge of continual learning in the open world lies precisely in *detecting* changes in the data distribution, which is why *effective OOD detection and lifelong learning go hand-in-hand*. In the future, I hope to extend this approach to an ensemble-based paradigm where multiple model states are updated during training and used during inference.

Coupling OOD detection with expansion-based continual learning is a simple and effective strategy that allows most DNNs used today to adapt to changes in the environment. Nevertheless, there are several practical challenges hindering the release of lifelong learning systems, which I explore in Chapters 10 through 13. The first is a lack of unified validation standards and frameworks that allow the easy evaluation of state-of-the-art models in dynamic environments, a fact that we look to mitigate for medical image segmentation with our *Lifelong nnU-Net* project (González et al., 2022c). Another factor is the key role that physicians must play in updating the database, which signifies an additional workload. *Active learning* can help reduce the amount of work, for which we propose the *i3Deep* method that allows the correction of segmentation masks with a refinement network trained prior to the interaction with the user (Gotkowski et al., 2022).

Finally, the most significant hurdle is the fact that, as of now, *systems that adapt post-deployment cannot be cleared for commercial use*. As I mentioned previously, and I describe in more detail in Chapter 13, there are several initiatives trying to amend this situation.

Yet considering the speed at which medical software devices are being developed *and authorized despite insufficient prospective validation* (van Leeuwen et al., 2021), coupled with the concerning lack of auditors who are allowed to provide certification in the EU (European Commission, 2023); it is paramount that we move to a better regulatory framework *soon*. Only with a swift change in regulation will we have products that truly remain safe and effective over time, and do not disappoint their users.

---

## List of Figures

---

2.1. Three mechanisms for automatic quality assurance . . . . .	16
2.2. Input space deemed acceptable by different quality assurance mechanisms . . . . .	17
3.1. Visual representation of four strategies for OOD detection . . . . .	18
6.1. Transferring an image to the training domain for downstream segmentation . . . . .	107
7.1. Continual learning scenarios . . . . .	111
7.2. Multi-head continual learning architecture . . . . .	113
7.3. Evaluation metrics for continual learning . . . . .	114
7.4. Commonly used continual learning strategies . . . . .	115
12.1. Interaction of a radiologist with an active learning system . . . . .	169
12.2. Additional efforts required from medical professionals . . . . .	189
13.1. Risks and opportunities of continual ML . . . . .	195
13.2. Possible framework for regulating continual ML products . . . . .	198
13.3. Advantages and disadvantages of locked vs. continual learning systems . . . . .	199

---

# Acknowledgements

---

I would like to thank all those who were my students at some point, including John Kalkhof, Karol Gotkowski, Hendrik Mehrrens, Marius Memmel, Antoine Sanner, Amin Ranem, Christian Harder, and Nick Lemke. Thank you for trusting me with your time and effort; it was an absolute pleasure to work with you. I would also like to thank Moritz Fuchs and Henry Krumb, who accompanied me these past three years and who reviewed virtually all my papers, giving me priceless feedback. Thank you also to Arjan Kuijper and Yannik Frisch for reviewing parts of this thesis, and, of course, Georgia Agelopoulou for her invaluable help.

I thank as well Dieter Fellner and Tianming Liu for serving as assessors on my doctoral committee, as well as Stefan Roth, Kristian Kersting and Reiner Hähnle for agreeing to sit on my examination board.

In addition, I thank Daniel Pinto dos Santos, Andreas Bucher and Ahmed Othman for their continued support and their willingness to invest time in advising me from their medical background. I am also grateful to all the fantastic colleagues I met through conferences or collaborations, many of whom have become my friends.

I thank as well the developers of the *nnU-Net* ([github.com/MIC-DKFZ/nnUNet](https://github.com/MIC-DKFZ/nnUNet)) and *TorchIO* ([torchio.readthedocs.io](https://torchio.readthedocs.io)) projects, which I used extensively in my work, particularly Fernando Pérez-García and Fabian Isensee. I am also extremely grateful to *ptrblck*, who I never met personally but who seems to have answered every question about *PyTorch* I could ever have.

Thank you also to my parents for giving me the freedom to pursue my own path and their unconditional support, and to my wonderful friends who always put up with my frequent disappearances near deadlines and join me for coffee or dinner when I need a break.

But most of all, thank you to Anirban Mukhopadhyay, without whom I *really* could not have written this thesis. Thank you for teaching me so much that I do not know who I would be if I had not been your student.

---

## Bibliography

---

- R. Aljundi, P. Chakravarty, and T. Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375, 2017.
- R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.
- R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia. Online continual learning with maximal interfered retrieval. *NeurIPS*, 32, 2019a.
- R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio. Gradient based sample selection for online continual learning. *NeurIPS*, 32, 2019b.
- A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.
- R. Alvarado. Should we replace radiologists with deep learning? pigeons, error and trust in medical ai. *Bioethics*, 36(2):121–133, 2022.
- R. Anzarouth, M. Muffoletto, I. Valasakis, and R. Sahar. Computer vision news - august 2021 - best of midl, 2021. URL <https://www.rsipvision.com/ComputerVisionNews-2021August/24/>. Accessed: 2023-02-01.
- N. Araslanov and S. Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021.
- N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.
- Y. Asano, C. Rupprecht, and A. Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2019.
- A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2019.
- M. E. Azzouzi, A. Mukhopadhyay, and H. Krumb. Medical device regulation of ai samd. NPR, 12 2022. URL <https://anchor.fm/anirban-mukhopadhyay7/episodes/Monir-El-Azzouzi-Medical-Device-Regulation-of-AI-SaMD-e1sjid7>.

- 
- W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 541–549. Springer, 2019.
- C. Baweja, B. Glocker, and K. Kamnitsas. Towards continual learning in medical imaging. *arXiv preprint arXiv:1811.02496*, 2018.
- E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.
- S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.
- P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *German Conference on Pattern Recognition*, pages 33–47. Springer, 2019.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical image analysis*, 58:101535, 2019.
- C. Chen, Q. Dou, H. Chen, and P.-A. Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*, pages 143–151. Springer, 2018.
- X. Chen, K. Men, B. Chen, Y. Tang, T. Zhang, S. Wang, Y. Li, and J. Dai. Cnn-based quality assurance for automatic segmentation of breast cancer in radiotherapy. *Frontiers in Oncology*, 10:524, 2020.
- Z. Chen and B. Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.
- L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni. Don’t forget, there is more than forgetting: new metrics for continual learning. In *Workshop on Continual Learning, NeurIPS 2018 (Neural Information Processing Systems)*, 2018.
- N. K. Dinsdale, M. Jenkinson, and A. I. Namburete. Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage*, 228:117689, 2021.
- A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

- 
- T. J. Draelos, N. E. Miner, C. C. Lamb, J. A. Cox, C. M. Vineyard, K. D. Carlson, W. M. Severa, C. D. James, and J. B. Aimone. Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 526–533. IEEE, 2017.
- S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach. Adversarial continual learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 386–402. Springer, 2020.
- K. O. Ellefsen, J.-B. Mouret, and J. Clune. Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLoS computational biology*, 11(4):e1004128, 2015.
- A. Elskhawy, A. Lisowska, M. Keicher, J. Henry, P. Thomson, and N. Navab. Continual class incremental learning for ct thoracic segmentation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 106–116. Springer, 2020.
- European Commission. European medical device nomenclature. <https://webgate.ec.europa.eu/dyna2/emdn/Z1206>, 2017a. Accessed: 2023-01-27.
- European Commission. Regulation of the european parliament and the council on medical devices, 2017b. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745&from=DE>.
- European Commission. Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>. Accessed: 2023-01-25.
- European Commission. Proposal for a regulation of the european parliament and the council amending regulations (eu) 2017/745 and (eu) 2017/746 as regards the transitional provisions for certain medical devices and in vitro diagnostic medical devices, 2023. URL [https://health.ec.europa.eu/system/files/2023-01/mdr\\_proposal.pdf](https://health.ec.europa.eu/system/files/2023-01/mdr_proposal.pdf). Accessed: 2023-01-27.
- European Council. Artificial intelligence act: Council calls for promoting safe ai that respects fundamental rights. [www.consilium.europa.eu](http://www.consilium.europa.eu), 2023. Accessed: 2023-01-25.
- A. O. Everhart, S. Sen, A. D. Stern, Y. Zhu, and P. Karaca-Mandic. Association between regulatory submission characteristics and recalls of medical devices receiving 510 (k) clearance. *JAMA*, 329(2): 144–156, 2023.
- A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDA 2020 and IKE 2020*, pages 877–894, 2021.
- FDA. Artificial intelligence and machine learning (ai/ml)-enabled medical devices. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>, 2023a. Accessed: 2023-01-27.
- FDA. Artificial intelligence and machine learning in software as a medical device. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>, 2023b. Accessed: 2023-01-27.



- 
- FDA. Feedback on the proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). <https://www.regulations.gov/document/FDA-2019-N-1185-0001>, 2023c. Accessed: 2023-01-25.
- FDA et al. Guidance for industry and fda staff: Modifications to devices subject to premarket approval (pma) - the pma supplement decision-making process, 2008. URL <https://www.fda.gov/media/73328/download>.
- FDA et al. Guidance for industry and food and drug administration staff—factors to consider when making benefit-risk determinations in medical device premarket approvals and de novo classifications. *Silver Spring: Center for Devices and Radiological Health, Center for Biologics Evaluation and Research, Food and Drug Administration*. Retrieved November, 3:2017, 2012.
- FDA et al. Deciding when to submit a 510 (k) for a software change to an existing device. guidance for industry and food and drug administration staff, 2017. URL <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-software-change-existing-device>.
- FDA et al. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). 2019. URL <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf>.
- FDA et al. Artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd) action plan. *US Food Drug Admin., White Oak, MD, USA, Tech. Rep*, 145022, 2021. URL <https://www.fda.gov/media/145022/download>.
- FDA et al. The software precertification (pre-cert) pilot program: Tailored total product lifecycle approaches and key findings. <https://www.fda.gov/media/161815/download>, 2022. Accessed: 2023-01-25.
- FDA et al. Digital health center of excellence. <https://www.fda.gov/medical-devices/digital-health-center-excellence>, 2023. Accessed: 2023-01-25.
- S. Fort, J. Ren, and B. Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.
- M. Fuchs, C. González, and A. Mukhopadhyay. Practical uncertainty quantification for brain tumor segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 407–422. PMLR, 2022.
- Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- I. Golan and R. El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.
- S. Golkar, M. Kagan, and K. Cho. Continual learning via neural pruning. *arXiv preprint arXiv:1903.04476*, 2019.

- 
- C. González and A. Mukhopadhyay. Self-supervised out-of-distribution detection for cardiac cmr segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 205–218. PMLR, 2021.
- C. González, K. Gotkowski, A. Bucher, R. Fischbach, I. Kaltenborn, and A. Mukhopadhyay. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 304–314. Springer, 2021.
- C. González, K. Gotkowski, M. Fuchs, A. Bucher, A. Dadras, R. Fischbach, I. J. Kaltenborn, and A. Mukhopadhyay. Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. *Medical Image Analysis*, 82:102596, 2022a.
- C. González, C. L. Harder, A. Ranem, R. Fischbach, I. J. Kaltenborn, A. Dadras, A. M. Bucher, and A. Mukhopadhyay. Quality monitoring of federated covid-19 lesion segmentation. In *Bildverarbeitung für die Medizin 2022*, pages 38–43. Springer, 2022b.
- C. González, A. Ranem, D. P. dos Santos, A. Othman, and A. Mukhopadhyay. Lifelong nnu-net: a framework for standardized medical continual learning. 2022c.
- C. González, A. Ranem, A. Othman, and A. Mukhopadhyay. Task-agnostic continual hippocampus segmentation for smooth population shifts. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 108–118. Springer, 2022d.
- C. González, N. Lemke, G. Sakas, and A. Mukhopadhyay. What is wrong with continual learning in medical image segmentation? *arXiv preprint arXiv:2010.11008*, 2023.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- B. Goodrich and I. Arel. Unsupervised neuron selection for mitigating catastrophic forgetting in neural networks. In *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 997–1000. IEEE, 2014.
- K. Gotkowski. 3D Deep Interactive Segmentation Based on Pre-acquired Uncertain 2D Patches. Master’s thesis, Technical University of Darmstadt, Germany, 2021.
- K. Gotkowski, C. González, A. Bucher, and A. Mukhopadhyay. M3d-cam: A pytorch library to generate 3d attention maps for medical deep learning. In *Bildverarbeitung für die Medizin 2021*, pages 217–222. Springer, 2021.
- K. Gotkowski, C. González, I. Kaltenborn, R. Fischbach, A. Bucher, and A. Mukhopadhyay. i3deep: Efficient 3d interactive segmentation with the nnu-net. In *International Conference on Medical Imaging with Deep Learning*, pages 441–456. PMLR, 2022.
- Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.
- R. Gu, J. Zhang, G. Wang, W. Lei, T. Song, X. Zhang, K. Li, and S. Zhang. Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures. *IEEE Transactions on Medical Imaging*, 42(1):245–256, 2022.
- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

- 
- A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.
- M. Hein, M. Andriushchenko, and J. Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.
- D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.
- D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.
- D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32:15663–15674, 2019.
- D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.
- T. R. Hoens, R. Polikar, and N. V. Chawla. Learning from streaming data with concept drift and imbalance: an overview. *Progress in Artificial Intelligence*, 1:89–101, 2012.
- J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):1–13, 2020.
- Y.-C. Hsu, Y.-C. Liu, A. Ramasamy, and Z. Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.
- Y. Hu, J. Jacob, G. J. Parker, D. J. Hawkes, J. R. Hurst, and D. Stoyanov. The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. *Nature Machine Intelligence*, 2(6):298–300, 2020.
- T. J. Hwang, A. S. Kesselheim, and K. N. Vokinger. Lifecycle regulation of artificial intelligence–and machine learning–based software devices in medicine. *Jama*, 322(23):2285–2286, 2019.
- IMDRF. Software as a medical device (samd): Key definitions, 2013. URL <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf>.
- F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- X. Jin, A. Sadhu, J. Du, and X. Ren. Gradient-based editing of memory examples for online task-free continual learning. *NeurIPS*, 34, 2021.
- A. Jungo and M. Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019.

- 
- A. Jungo, F. Balsiger, and M. Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.
- J. Kalkhof, C. González, and A. Mukhopadhyay. Disentanglement enables cross-domain hippocampus segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.
- N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu. A lifelong learning approach to brain mr segmentation across scanners and protocols. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–484. Springer, 2018.
- D. Karimi, S. D. Vasylechko, and A. Gholipour. Convolution-free medical image segmentation using transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 78–88. Springer, 2021.
- C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9, 2019.
- B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, and T. Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157. IEEE, 2019.
- J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- A. Knoblauch, E. Körner, U. Körner, and F. T. Sommer. Structural synaptic plasticity has high memory capacity and can explain graded amnesia, catastrophic forgetting, and the spacing effect. *PloS one*, 9(5):e96485, 2014.
- S. A. Kohl, B. Romera-Paredes, C. Meyer, J. D. Fauw, J. R. Ledsam, K. H. Maier-Hein, S. A. Eslami, D. J. Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6965–6975, 2018.
- A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). 2020. URL <http://www.cs.toronto.edu/~kriz/cifar.html>.
- D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.
- N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 7–12. IEEE, 2020.
- B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30:6402–6413, 2017.
- H. H. Lee, Y. Tang, O. Tang, Y. Xu, Y. Chen, D. Gao, S. Han, R. Gao, M. R. Savona, R. G. Abramson, et al. Semi-supervised multi-organ segmentation through quality assurance supervision. In *Medical Imaging 2020: Image Processing*, volume 11313, pages 363–369. SPIE, 2020.

- 
- K. Lee, H. Lee, K. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018a.
- K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018b.
- K. Lee, K. Lee, J. Shin, and H. Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 312–321, 2019.
- S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, and D. Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.
- M. Lenga, H. Schulz, and A. Saalbach. Continual learning for domain adaptation in chest x-ray classification. In *International Conference on Medical Imaging with Deep Learning*, 2020.
- X. Li, C. Desrosiers, and X. Liu. Symmetric contrastive loss for out-of-distribution skin lesion detection. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.
- Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.
- T. Liu, E. Siegel, and D. Shen. Deep learning and medical image analysis for covid-19 diagnosis and prediction. *Annual Review of Biomedical Engineering*, 24:179–201, 2022a.
- W. Liu, X. Wang, J. D. Owens, and Y. Li. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*, 2020.
- X. Liu, B. Glocker, M. M. McCradden, M. Ghassemi, A. K. Denniston, and L. Oakden-Rayner. The medical algorithmic audit. *The Lancet Digital Health*, 2022b.
- J. Makower, A. Meer, and L. Denend. Fda impact on us medical technology innovation: a survey of over 200 medical technology companies. *Arlington (Virginia): National Venture Capital Association*, 2010.
- A. Mallya and S. Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.
- Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.
- S. Medtech. Press release 28.11.22 - politicians decide in favour of patient care. <https://www.swiss-medtech.ch/en/news/politicians-decide-favour-patient-care>, 2023. Accessed: 2023-01-27.
- A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.
- H. A. Mehrtens. Leveraging functional diversity in ensembles to improve uncertainty estimation and calibration. Master’s thesis, Technical University of Darmstadt, Germany, 2021.

- 
- H. A. Mehrrens, C. González, and A. Mukhopadhyay. Improving robustness and calibration in ensembles with diversity regularization. In *DAGM German Conference on Pattern Recognition*, pages 36–50. Springer, 2022.
- M. Memmel, C. González, and A. Mukhopadhyay. Adversarial continual learning for multi-domain hippocampal segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 35–45. Springer, 2021.
- U. Michieli and P. Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1114–1124, 2021.
- A. Midya, J. Chakraborty, M. Gönen, R. K. Do, and A. L. Simpson. Influence of ct acquisition and reconstruction parameters on radiomic feature reproducibility. *Journal of Medical Imaging*, 5(1): 011020, 2018.
- S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, pages 5216–5223, 2020.
- M. Monteiro, L. Le Folgoc, D. Coelho de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12756–12767. Curran Associates, Inc., 2020.
- U. J. Muehlematter, P. Daniore, and K. N. Vokinger. Approval of artificial intelligence and machine learning-based medical devices in the usa and europe (2015–20): a comparative analysis. *The Lancet Digital Health*, 3(3):e195–e203, 2021.
- M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, and M. Maruthappu. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *bmj*, 368, 2020.
- Open Regulatory. Notified body capacity and reviews. <https://openregulatory.com/notified-bodies/>, 2023. Accessed: 2023-01-25.
- O. Ostapenko, M. Puscas, T. Klein, P. Jahnichen, and M. Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019.
- Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.
- F. Ozdemir, P. Fuernstahl, and O. Goksel. Learn the new, keep the old: Extending pretrained models with new anatomy and images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 361–369. Springer, 2018.
- S. Özgün, A.-M. Rickmann, A. G. Roy, and C. Wachinger. Importance driven continual learning for segmentation across domains. In *International Workshop on Machine Learning in Medical Imaging*, pages 423–433. Springer, 2020.

- 
- T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu. Improving adversarial robustness via promoting ensemble diversity. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4970–4979. PMLR, 09–15 Jun 2019.
- M. Parekh, A. Donuru, R. Balasubramanya, and S. Kapur. Review of the chest ct differential diagnosis of ground-glass opacities in the covid era. *Radiology*, 297(3):E289–E302, 2020.
- S. Pati, S. P. Thakur, M. Bhalerao, S. Thermos, U. Baid, K. Gotkowski, C. González, O. Guley, I. E. Hamamci, S. Er, et al. Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*, 2021.
- A. Patra and J. A. Noble. Incremental learning of fetal heart anatomies using interpretable saliency maps. In *Medical Image Understanding and Analysis: 23rd Conference, MIUA 2019, Liverpool, UK, July 24–26, 2019, Proceedings 23*, pages 129–141. Springer, 2020.
- M. Perkonigg, J. Hofmanninger, C. J. Herold, J. A. Brink, O. Pinykh, H. Prosch, and G. Langs. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nature Communications*, 12(1):1–12, 2021.
- J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.
- S. Pidhorskyi, R. Almoheisen, and G. Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in neural information processing systems*, pages 6822–6833, 2018.
- J. Qi, K. Tang, Q. Sun, X.-S. Hua, and H. Zhang. Class is invariant to context and vice versa: on learning invariance for out-of-distribution generalization. In *European Conference on Computer Vision*, pages 92–109. Springer, 2022.
- A. Ranem, C. González, and A. Mukhopadhyay. Continual hippocampus segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3711–3720, 2022.
- D. Rao, F. Visin, A. Rusu, R. Pascanu, Y. W. Teh, and R. Hadsell. Continual unsupervised representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- R. Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.
- H. Ravishankar, R. Venkataramani, S. Anamandra, P. Sudhakar, and P. Annangi. Feature transformers: privacy preserving lifelong learners for medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 347–355. Springer, 2019.
- S.-A. Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516, 2017a.
- S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017b.

- 
- M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3): 199–217, 2021.
- M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3379–3388, 2018.
- A. Sanner, C. González, and A. Mukhopadhyay. How reliable are out-of-distribution generalization methods for medical image segmentation? In *DAGM German Conference on Pattern Recognition*, pages 604–617. Springer, 2021.
- P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.
- R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.
- H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2017.
- D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.
- A. Smith and M. Severn. An overview of continuous learning artificial intelligence-enabled medical devices. *Canadian Journal of Health Technologies*, 2(5), 2022.
- S. Srivastava, M. Yaqub, K. Nandakumar, Z. Ge, and D. Mahapatra. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 226–238. Springer, 2021.
- A. V. Terekhov, G. Montone, and J. K. O’Regan. Knowledge transfer in deep block-modular neural networks. In *Conference on Biomimetic and Biohybrid Systems*, pages 268–279. Springer, 2015.
- L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.
- V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*, 36(8):1597–1606, 2017.
- G. M. van de Ven, T. Tuytelaars, and A. S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.



- 
- K. van Garderen, S. van der Voort, F. Incekara, M. Smits, and S. Klein. Towards continuous learning for glioma segmentation with elastic weight consolidation. *MIDL*, 2019.
- K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, and M. de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European radiology*, 31(6):3797–3804, 2021.
- K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, and M. de Rooij. Ai for radiology. <https://grand-challenge.org/aiforradiology/>, 2023. Accessed: 2023-01-27.
- E. M. van Rikxoort, B. de Hoop, M. A. Viergever, M. Prokop, and B. van Ginneken. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical physics*, 36(7):2934–2947, 2009.
- J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler. Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights into imaging*, 11(1):1–16, 2020.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- R. Venkataramani, H. Ravishankar, and S. Anamandra. Towards continuous domain adaptation for medical imaging. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 443–446. IEEE, 2019.
- K. N. Vokinger and U. Gasser. Regulating ai in medicine in the united states and europe. *Nature Machine Intelligence*, 3(9):738–739, 2021.
- K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim. Continual learning in medical devices: Fda’s action plan and beyond. *The Lancet Digital Health*, 3(6):e337–e338, 2021.
- A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 550–564, 2018.
- H. Wang, A. Zhang, Y. Zhu, S. Zheng, M. Li, A. J. Smola, and Z. Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pages 23446–23458. PMLR, 2022.
- Y.-X. Wang, D. Ramanan, and M. Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2471–2480, 2017.
- Y. Wen, D. Tran, and J. Ba. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. In *Eighth International Conference on Learning Representations (ICLR 2020)*, 2020.
- J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.
- E. Wu, K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27(4):582–584, 2021.
- M. Wu and N. Goodman. A simple framework for uncertainty in contrastive learning. *arXiv preprint arXiv:2010.02038*, 2020.

- 
- Y. Xia, Y. Zhang, F. Liu, W. Shen, and A. L. Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020.
- W. Yan, Y. Wang, S. Gu, L. Huang, F. Yan, L. Xia, and Q. Tao. The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–631. Springer, 2019.
- J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 255–263. Springer, 2019.
- J. Yoon, J. Lee, E. Yang, and S. J. Hwang. Lifelong learning with dynamically expandable network. In *International Conference on Learning Representations*. International Conference on Learning Representations, 2018.
- N. Yu, L. S. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019.
- J. Zhang, R. Gu, G. Wang, and L. Gu. Comprehensive importance-based selective regularization for continual segmentation across multiple sites. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 389–399. Springer, 2021.
- J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.