Computer Aided Medical Procedures
Prof. Dr. Nassir Navab

Dissertation

Learning Robust Representations for Medical Diagnosis

Magdalini Paschali

Fakultät für Informatik
Technische Universität München

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin

Learning Robust Representations for Medical Diagnosis

Magdalini Paschali

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r): Prof. Dr. Julien Gagneur

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab

2. Prof. Dr. Daniel Rückert

Die Dissertation wurde am 11.05.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 06.09.2021 angenommen.

Magdalini Paschali
Learning Robust Representations for Medical Diagnosis
Dissertation, Version 2.0

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 and Garching bei München

Abstract

Medical tasks such as patient diagnosis, treatment planning, and surgery are challenging not only for clinicians but also for computer-aided systems. Such systems powered by machine learning could significantly aid physicians with these tasks and greatly impact the future of healthcare. However, training such systems requires a significant amount of training data with accompanying expert annotations. Furthermore, challenges such as outliers, rare medical conditions, image artefacts, site variations, and inter-observer variability lead to an additional burden to computer-aided systems' generalization ability and robustness. Moreover, threats like adversarial attacks could pose a security risk for healthcare with serious implications.

To this end, this dissertation tackles the issues of improving and thoroughly evaluating the robustness of machine learning models for medical diagnosis.

The first part describes two methods to improve model robustness for medical image classification and segmentation. A novel data augmentation technique is proposed that utilizes manifold-exploring geometric transformations. Our method improves model robustness against affine and projective transformations and increases model performance on fine-grained skin lesion and breast tumor classification. A metric based on geodesic distance is introduced to quantify the robustness of classifiers by measuring the distance to their decision boundary. Finally, a ternary quantization method is described, that in addition to compressing the size of a trained model by 16 times, enhances the training dynamics of a large volumetric model for whole-brain segmentation.

In the second part, we introduce methods to evaluate the robustness of classifiers. Our novel benchmarking strategy utilizes adversarial examples to evaluate various deep learning models for classification and segmentation. Our method highlights that models that achieve similar or identical performance on clean test data could have substantial differences regarding robustness to adversarial attacks.

Finally, we elaborate on robustness beyond imaging data and present a novel analysis pipeline for depression score prediction in adolescents utilizing neuropsychological and clinical data. Our pipeline consists of a longitudinal, multi-task model that accurately predicts depression scores, a permutation scheme that identifies significant feature categories in the neuropsychological and clinical assessments, and model interpretation that ranks the importance of each feature.

Zusammenfassung

Medizinische Aufgaben wie Patientendiagnose, Behandlungsplanung und Chirurgie sind nicht nur für Kliniker, sondern auch für computergestützte Systeme eine Herausforderung. Solche Systeme, die auf maschinellem Lernen basieren, könnten Ärzten bei diesen Aufgaben erheblich helfen und die Zukunft des Gesundheitswesens stark beeinflussen. Das Training solcher Systeme erfordert jedoch eine erhebliche Menge an Trainingsdaten mit begleitenden Expertenkommentaren. Darüber hinaus führen Herausforderungen wie Ausreißer, seltene medizinische Bedingungen, Bildartefakte, Standortvariationen und die Variabilität zwischen den Beobachtern zu einer zusätzlichen Belastung für die Generalisierungsfähigkeit und Robustheit computergestützter Systeme. Darüber hinaus können Bedrohungen wie adversarische Angriffe ein Sicherheitsrisiko für das Gesundheitswesen mit schwerwiegenden Folgen darstellen.

Aus diesem Grund befasst sich diese Dissertation mit der Verbesserung und gründlichen Evaluierung der Robustheit von maschinellen Lernmodellen für die medizinische Diagnose.

Im ersten Teil werden zwei Methoden zur Verbesserung der Modellrobustheit für die Klassifikation und Segmentierung medizinischer Bilder beschrieben. Es wird eine neuartige Technik zur Datenerweiterung vorgeschlagen, die geometrische Transformationen zur Erforschung von Mannigfaltigkeiten nutzt. Diese Methode verbessert die Robustheit des Modells gegenüber affinen und projektiven Transformationen und erhöht die Leistung des Modells bei der Klassifizierung von feinkörnigen Hautläsionen und Brusttumoren. Eine Metrik, die auf der geodätischen Distanz basiert, wird eingeführt, um die Robustheit von Klassifikatoren zu quantifizieren, indem die Distanz zu ihrer Entscheidungsgrenze gemessen wird. Schließlich wird eine ternäre Quantisierungsmethode beschrieben, die nicht nur die Größe eines trainierten Modells um das 16-fache komprimiert, sondern auch die Trainingsdynamik eines großen volumetrischen Modells für die Segmentierung des gesamten Gehirns verbessert.

Im zweiten Teil stellen wir Methoden vor, um die Robustheit von Klassifikatoren zu bewerten. Unsere neuartige Benchmarking-Strategie nutzt adversarische Beispiele, um verschiedene Deep-Learning-Modelle für Klassifizierung und Segmentierung zu bewerten. Unsere Methode verdeutlicht, dass Modelle, die eine ähnliche oder identische Leistung auf sauberen Testdaten erzielen, erhebliche Unterschiede in Bezug auf die Robustheit gegenüber gegnerischen Angriffen aufweisen können.

Abschließend gehen wir auf die Robustheit jenseits von Bildgebungsdaten ein und präsentieren eine neuartige Analyse-Pipeline zur Vorhersage von Depressionswerten bei Jugendlichen unter Verwendung neuropsychologischer und klinischer Daten. Unsere Pipeline besteht aus einem longitudinalen Multi-Task-Modell, das Depressionswerte genau vorhersagt, einem Permutationsschema, das signifikante Merkmalskategorien in den neuropsychologischen und klinischen Bewertungen identifiziert, und einer Modellinterpretation, die die Wichtigkeit jedes Merkmals einstuft.

Acknowledgments

I would first like to thank my advisor Prof. Nassir Navab who trusted me with this opportunity and gave me the freedom to explore and learn over the past four years. He was a great source of advice and support and gave me the chance to travel the world, meet amazing people and grow in such a supportive, warm environment like CAMP.

I am also very grateful for the support of Thomas Wendler, who helped transform the lab, created a great collaborative atmosphere, and gave me great advice. Also, big thanks to Martina Hilla for taking care of everything.

I would also like to thank the Computational Neuroscience group at Stanford, Prof. Kilian Pohl, Ehsan Adeli, Quinyu Zhao, Orsolya Kiss, and Fiona Baker. Having a chance to collaborate with such an amazing team, even remotely, was exactly what I needed after a tough 2020 to motivate me, teach me many new things, and introduce me to a research direction that I am very excited to follow!

Many thanks go to Abhijit Guha Roy, who guided me, especially at the beginning of my Ph.D., and helped me find my direction. Thanks to Sailesh Conjeti for teaching me how to write a paper and preparing me for this journey with good advice.

I am very grateful for the amazing IFL team and would like to thank Farid Azampour, Matthias Keicher, Christine Eilers and Maria Tirindelli, who are more friends than colleagues and taught me so much along the way. Special thanks go to Tobias Czempiel for his help, support, and patience.

Thank you to my oldest friends Greg and Chris, for making the past ten years and my whole student life so much more fun.

I would also like to thank my parents Paschalis and Daphne, for believing in me and supporting me all those years. Many thanks to my brother Sotiris, who was always ready to help me with his creativity and excitement.

Finally, thanks to Walter Simson for his love, acceptance, and constant encouragement. His belief in me helped me embrace my mistakes, step out of my comfort zone and enjoy this journey every day.

Contents

Part I

Introduction

Introduction 1

1.1 Learning Robust Representations

1.1.1 Introduction

Machine learning systems have been integrated into commonly-used systems that provide movie and music recommendations, personalized advertisements, cooking assistance, autonomous transportation, and medical diagnosis [223]. Such systems can rely on statistical methods like linear regression or be powered by more powerful models, such as Deep Neural Networks (DNNs) [104].

The study and deployment of DNNs has been a drastically growing field in both research and industry and includes developing new model architectures and modules that provide improved trainability, faster and more stable optimization algorithms, larger and more diverse datasets, and more.

With the increasing popularity and deployment of DNNs in applications, such as financial systems, autonomous driving, and healthcare, the security and privacy of DNNs became an active area of investigation. Security breaches could occur during training or testing of a model, or even afterwards in the form of model or dataset theft [128].

A particularly interesting direction regarding security is adversarial inputs, which were first formally described by Biggio and Roli [29] as a way to circumvent spam filters. Since then, Szegedy et al. [273] introduced these inputs on imaging data as adversarial examples. Samples, crafted with the intention to fool machine learning models without bearing any detectable distortion by the human eye when compared to their benign sources.

However, *adversarial examples* are only one of the many adverse types of inputs that could challenge the predictions of a DNN. *Outliers*, data that differ significantly from the observations a model was trained on could also create problems in the performance of a machine learning model. Furthermore, *unseen or rare findings*, commonly observed in medical imaging as rare diseases or anatomical variations are another source of performance deterioration. Variations in perspective, illumination changes, and transformations such as rotations and translations are additional types of input variabilities that could lead to model failure.

1.1.2 Generalizability *vs.* Robustness

In order to deploy models in real-world applications successfully, it would be crucial to ensure that they are characterized by the following attributes:

- **Generalizability**: The ability of models to generalize to unseen data, originating from the same distribution as the training data. This property is closely related to model overfitting and data memorization, both of which lead to a significant drop in model performance in the real-world [307].

- **Robustness**: The ability of a model to maintain acceptable performance when tested on data from a distribution different than the one of the training data [122]. As discussed above, there are various types of adverse inputs. Therefore, a machine learning-powered application must define what kind of input data it aims to be robust to in advance.

This thesis first describes methods to improve a model's robustness [217] and to enhance a model's training dynamics, which can lead to better generalizability [216]. Afterwards, an evaluation method for the robustness of DNNs is introduced [215]. Finally, a pipeline for the analysis of a model is discussed, focusing on non-imaging longitudinal data.

1.1.3 Robustness in Medical Diagnosis

Medical diagnostics systems powered by machine learning are reported to have achieved similar performance with physicians on applications in radiology, pathology, dermatology, and ophthalmology, to name a few [61]. Many aspects of healthcare can be aided by machine learning systems, including diagnosis, treatment planning, surgery, patient triage, and health insurance claim approvals [288].

Analyzing medical information is executed using standardized Medical Imaging formats like Digital Imaging and Communications in Medicine (DICOM) [199]. Medical images are shared within a hospital's Picture Archiving and Communication Systems (PACS) to allow efficient archiving and data analysis. Moreover, Electronic Health Records (EHR) is a standardized collection of patient and population health information electronically stored in a digital format.

Health insurance providers, pharmaceutical companies, and healthcare providers constitute some of the stakeholders within the healthcare system with strong financial interests [239].

Furthermore, hospitals are underfunded in various parts of the world [1]; thus, their computer infrastructure systems could be severely outdated and lack technical personnel to update software and hardware to the latest security measures [135].

The above factors, namely the extensive digitization, the outdated infrastructure, and the high financial interests, constitute healthcare especially vulnerable to security threats. In recent years cyber attacks against systems using artificial intelligence leverage data poisoning, model theft, and adversarial examples [86]. Adversarial approaches could enable billing teams of insurance companies to imperceptibly alter insurance claims without getting detected by fraud detectors [239]. Manufacturing companies could also be tempted to utilize adversarial methods for the approval of drugs and devices. Pharmaceutical corporations could deploy adversarial approaches to maliciously bias trial outcomes [86].

Stites et al. [265] built an application that scans the World Wide Web and locates available and unprotected radiology servers. Their scan found 2774 radiology or DICOM servers that were unprotected worldwide. 719 of those servers were fully open to sharing patient data communications [265]. Furthermore, in the past 10 years, approximately 3000 breaches, each including more than 500 medical records, have taken place in the United States [35].

CyberAngel [203] created a framework that scanned around 4.3 billion IP addresses and found more than 45 million unique medical images on over 2,140 unprotected servers across countries, including the US, UK, France, and Germany. They also discovered that openly available medical images were accompanied by up to 200 lines of metadata per patient, which included information about the patient name, birth date, address, height, weight, diagnosis, and more [203].

In 2019, Mirsky et al. [199] showed how an attacker could leverage a deep learning system to insert or remove abnormal findings on CT and MRI scans in DICOMs from the scanner to the PACS. Two deep networks were used, one for injection and the other for removal of lung nodules. The altered images were so realistic that they fooled 99% of radiologists who evaluated them [75, 199].

Another type of malware that challenges organizations around the world is "ransomware". This is a malicious software that, once downloaded and executed, encrypts as many files as possible and then demands a ransom payment to recover the files [82]. Healthcare is one of the most affected fields, with 15% of ransomware globally found in healthcare institutions in 2017 [207].

Exposure to such sensitive information can lead to identity theft, fraud [239], and millions of dollars in litigation costs from lawsuits by law firms and victims of such breaches [75].

Investigating and evaluating the robustness of DNNs for medical applications could alleviate some of the security threads in healthcare and facilitate the resilience of AI systems.

2 Contributions

This thesis is built from the following contributions. In Part II, we first give an overview of adversarial attacks and defenses and their applications in medical imaging. Afterwards, we discuss methods to improve a model's robustness to geometric transformations. Finally, we introduce a ternary quantization method for large volumetric models that leads to improved training dynamics:

- **M. Paschali**, W. Simson, A. Guha Roy, R. Göbl, C. Wachinger, N. Navab. *"Manifold Exploring Data Augmentation with Geometric Transformations for Increased Performance and Robustness."* International Conference on Information Processing in Medical Imaging (IPMI), 2019

- **M. Paschali***, S. Gasperini*, A. Guha Roy, M.Y.-S. Fang, N. Navab. *"3DQ: Compact Quantized Neural Networks for Volumetric Whole Brain Segmentation."* International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Shenzhen, 2019 (Equal Contribution)

In Part III, we address robustness evaluation with an emphasis on standardized model benchmarking. We then introduce a novel model evaluation technique using adversarial examples proposed in the following contribution:

- **M. Paschali**, S. Conjeti, F. Navarro, N. Navab. *"Generalizability vs. Robustness: Adversarial Examples for Medical Imaging."* International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Granada, 2018

Finally, in Part IV, we tackle robustness beyond imaging data and present a novel pipeline that combines deep learning with statistical evaluation to perform longitudinal prediction from tabular data presented in:

- **M. Paschali**, O. Kiss, Q. Zhao, E. Adeli, S. Podhajsky, I. Gotlib, K.M. Pohl, F. Baker. *"Predicting Symptoms of Depression in Adolescents based on Longitudinal Self-Reports and Behavioral Assessments."*, 2021 (Under Review)

Part II

Robustness Improvement

Adversarial Examples 3

3.1 Introduction

As discussed in the previous chapter, robustness is a broad term that describes the resilience of a deep model to test samples that do not originate from the same distribution as the training data. Therefore, to improve the robustness of a model, we need to first specify what kind of robustness is critical to a particular application. Some methods, like data augmentation, are general and can benefit both the generalizability and the robustness of a model. Others like adversarial training are more specific to a use case, namely adversarial examples. Thus in this Part of the dissertation, we will first discuss ways to improve a model's robustness to adversarial attacks and then introduce two contributions, a data augmentation method tailored to increase robustness to geometric transformations and a ternary quantization mechanism that enhances training dynamics and model performance.

Szegedy et al. [273] discovered that minute changes to the input of a neural network can have significant effects on its output. Specifically, a minor perturbation of pixels in the input image to a classifier can completely change the class predicted by the model. Moreover, the difference between the original and perturbed examples is often imperceptible to the human eye. This sensitivity to small perturbations has been found to exist not only in neural networks but also in traditional machine learning techniques, like linear models and nearest neighbor classifiers [212]. Even though the human brain is not fooled in the same way as machine learning models, optical illusions, like the ones shown in Fig. 3.1 trick the human brain similarly to how adversarial examples fool classifiers.

Such minimally perturbed samples that are able to fool machine learning models are called *adversarial examples* and are an interesting area of research for various reasons. First, they highlight that machine learning methods do not yet fully understand the tasks they are trained

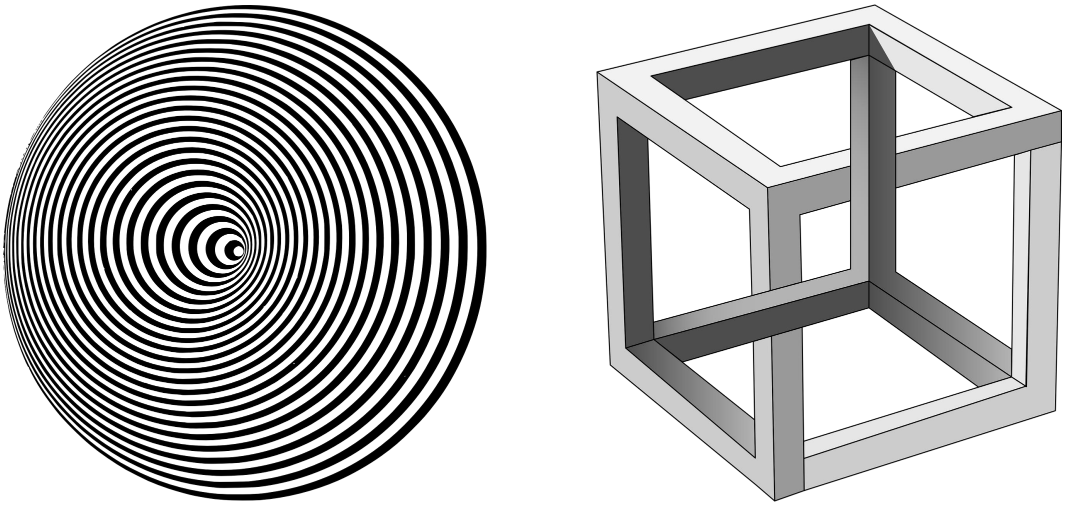


Fig. 3.1. Adversarial examples have similar effect to machine learning models like optical illusions to the human brain. Left: Concentric circles illusion. Right: The impossible cube or irrational cube, an impossible object invented by M.C. Escher [220].

to perform, even when these methods achieve human level performance on a test set consisting of natural inputs. Thus, if we improve the performance on adversarial examples, machine learning models will acquire a better understanding of underlying tasks and make the right decisions for the right reasons. Second, adversarial examples can have crucial implications for computer security, which is particularly interesting for healthcare as discussed earlier in the Introduction.

A noteworthy property of adversarial examples is that a specific adversarial example that was crafted to deceive one classifier, model A, will often also deceive another model, model B. When model B has a different architecture than model A, this phenomenon is called *cross-model generalization* of adversarial examples. Moreover, if model B was trained on a different training dataset than model A, this is called *cross-dataset generalization* [102, 116, 273, 282]

Fig. 3.2 shows the learned decision boundaries of models A and B and the actual task boundary between classes 1 and 2 [282]. Since the models boundaries are similar, an adversarial example generated for model A can also often cross the boundary of model B. The trasferability of adversarial examples shows that they pose a security threat even when the attacker does not have access to the target's model architecture, weights, or training set.

Based on their knowledge of the target, adversarial attacks can be split into *white-box attacks*, where the adversary has full access to the model and its parameters to generate adversarial examples and *black-box* attacks, where the adversary has no or limited knowledge about the model, and crafts adversarial examples without any gradient information using an independent classifier [212, 213].

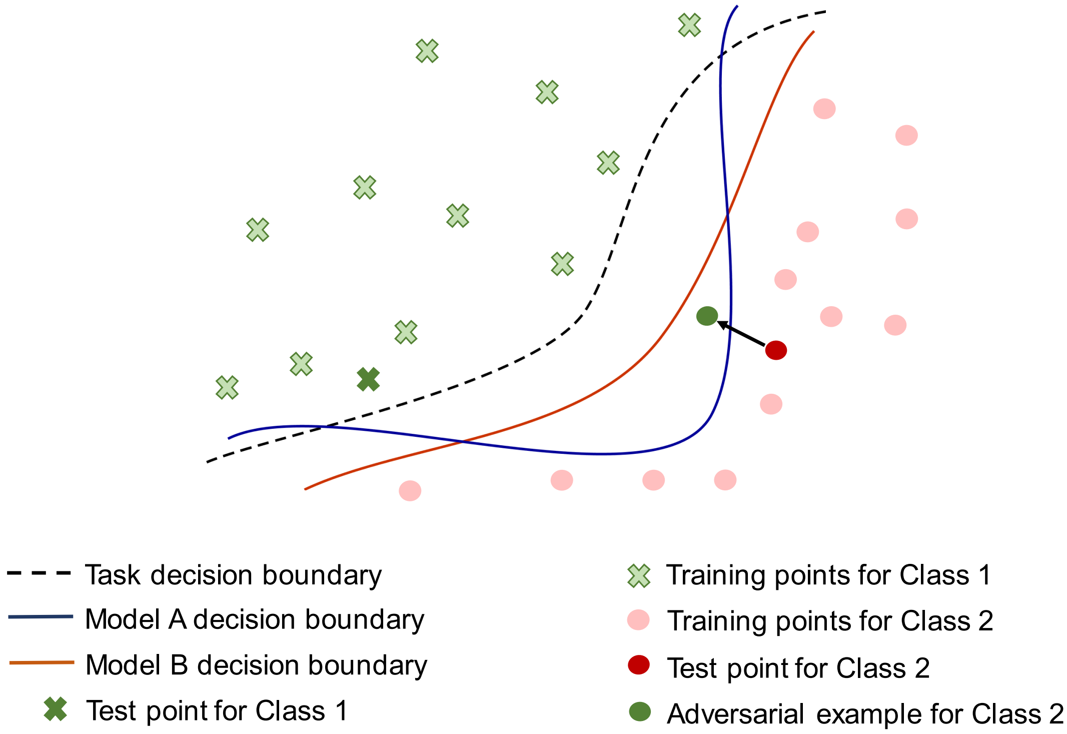

Task decision boundary
Model A decision boundary
Model B decision boundary
Test point for Class 1
Training points for Class 1
Training points for Class 2
Test point for Class 2
Adversarial example for Class 2

Fig. 3.2. Model decision boundaries and adversarial example transferability. Model boundaries are similar, hence an adversarial example generated for model A can also often cross the boundary of model B [282].

3.1.1 The Linearity Hypothesis

Initially the cause for the existence of adversarial examples was unknown [273]. First, they were attributed to the high complexity and non-linearity of neural networks. However, Goodfellow et al. [102] found that adversarial examples affect simple models, such as shallow linear classifiers in a similar way to deep models. Thus, they introduced the *linearity hypothesis*, which states that adversarial examples exist due to the fact that models behave extremely linearly as a function of their inputs.

This hypothesis is based on the fact that deep neural networks (DNNs) often use components that are extremely linear, such as rectified linear units (ReLUs) [100]. Even though, DNNs are nonlinear as a function of their parameters, they are linear as a function of their inputs. Networks using ReLUs divide input space into several regions, with the output of the rectified linear layers being linear within each region. To grasp the reason linear functions are vulnerable to adversarial examples, consider the output of a model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. If the input $\mathbf{x}$ is perturbed by $\epsilon \cdot \text{sign}(\mathbf{w})$, then the output increases by $\epsilon||\mathbf{w}||_1$. For a high dimensional $\mathbf{w}$, the increase in the output can be extremely large [102, 116].

Next we will discuss how adversarial examples are distributed in space. Initially, it was hypothesized that they were rare and occurred in small pockets around the decision boundaries of a classifier that can only be found with careful search strategies [273]. However, according to the linearity hypothesis adversarial examples take up large volumes of space. If a loss function $\mathcal{L}(\mathbf{x}, y)$ increases in a linear fashion towards a direction $\mathbf{d}$, then an adversarial example $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\epsilon}$ will be misclassified as long as $\boldsymbol{\epsilon}^T\mathbf{d}$ is large. Thus, the linearity hypothesis dictates that a hyperplane where $\boldsymbol{\epsilon}^T\mathbf{d} = C$ for a constant C divides the space into two half-spaces. The original input $\mathbf{x}$ along with a large region of samples on the same side of the hyperplane as $\mathbf{x}$ are correctly classified. However, on the opposite side of the hyperplane, nearly all points have a different classification [102, 116].

3.1.2 Types of Adversarial Attacks

The adversarial perturbations explained above are responsible for *evasion attacks* and require minimal changes to be done to the original test images. However, an attacker cannot always craft adversarial perturbations at test-time.

It has been shown [28] that machine learning systems are still susceptible to another type of attack called *poisoning attacks*. These attacks occur at training time; the attacker manipulates the performance of a model by inputting carefully constructed poisoned samples into the training data. In [247] those poison instances are constructed using a watermarking strategy, and the attacker is populating the training dataset with clean and poisoned samples to cause deterioration in the model performance.

Another type of data poising attacks are *backdoor attacks* [59]. A backdoor is a kind of input that the model's creator is unaware of but that the attacker can use to manipulate the ML system predictions. For instance, an attacker could teach a malware classifier that if a certain artefact is present in a file, the file should always be classified to a specific class. That way, the attacker can create an adversarial input as they insert that artefact somewhere into their file [221]. Additionally, it has been shown [106] that poisoned models can be transferred using transfer learning even on a completely different dataset.

Finally, model stealing or extraction describes an attack, where a black box machine learning system is interrogated to either reconstruct the model or extract the data it was trained on. This could lead to critical issues regarding personal private training data or confidential and sensitive models [162]. Krishna et al. [162] showed that an attacker could steal natural language processing models without access to any input training data. Their proposed attack inputs randomly-sampled sequences of words to a victim model and fine-tunes their own classifier on the labels predicted by the victim model [128].

Kaissis et al. [141] thoroughly discuss how such attacks to models or datasets can have a critical impact in healthcare and how differential privacy [80] can be applied to the input data, the results of an algorithm or the algorithm itself to provide resistance to such attacks.

Within the scope of this dissertation, we will focus on evasion attacks with adversarial examples and their defenses, methods to improve model training dynamics and robust evaluation pipelines for imaging and tabular data.

3.2 Crafting Adversarial Examples

Let $\mathbf{x} \in \mathbb{R}^n$ be a vector of input image features, such as pixels, and y be an integer specifying the classification label for $\mathbf{x}$. Let f be the classification function learned by the model, so that $f(\boldsymbol{x})$ is the prediction of the model. Let $\mathcal{L}(\mathbf{x}, y)$ be the cost used to train f [116]. The goal of adversarial example crafting is to find an input point $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon}$ that causes misclassification by f. Various methods for adversarial example crafting have been proposed using different criteria to determine the model's performance, different strategies to minimize the size of $\boldsymbol{\epsilon}$, and different approximations to optimize the selected criterion [102, 116, 273]. Since the goal of adversarial examples is to be imperceptible by the human eye, every adversarial crafting method aims at finding the minimum $\boldsymbol{\epsilon}$ that will still fool a trained model.

3.2.1 Gradient-based Attacks

Szegedy et al. [273] proposed the first adversarial crafting method, based on solving the following optimization problem:

$$\boldsymbol{\epsilon} = \text{argmin}_{\boldsymbol{\epsilon}} \lambda ||\boldsymbol{\epsilon}||_2^2 + \mathcal{L}(\boldsymbol{x} + \boldsymbol{\epsilon}, \hat{y}) \tag{3.1}$$

for $(\boldsymbol{x} + \boldsymbol{\epsilon} \in [0,1]^n)$ and where $\hat{y}$ is the target class the adversarial example should be misclassified to [273]. They used box-constrained L-BFGS to perform the minimization, which was repeated iteratively with multiple values of λ so that the minimum amount of perturbation $\boldsymbol{\epsilon}$ that can fool f can be found. The method is computationally expensive, due to the iterative optimization procedure for each example, however it is highly effective in crafting imperceptible adversarial examples with high success rate [116, 273].

Goodfellow et al. [102] simplified the problem and proposed the *Fast Gradient Sign Method* (FGSM). They calculated the added perturbation $\boldsymbol{\epsilon}$ as follows:

$$\boldsymbol{\epsilon} = \text{argmax}_{\boldsymbol{\epsilon}} \mathcal{L}(\boldsymbol{x} + \boldsymbol{\epsilon}, y) \tag{3.2}$$

In this case $||\boldsymbol{\epsilon}||_\infty < \eta$, where η is a hyperparameter chosen by the attacker to specify the maximum allowed amount of perturbation to the input image pixels. To obtain a fast, closed-form solution, Goodfellow et al. [102] replaced $\mathcal{L}$ with a first-order Taylor series approximation:

$$\boldsymbol{\epsilon} = \text{argmax}_{\boldsymbol{\epsilon}} \mathcal{L}(\boldsymbol{x}, y) + \boldsymbol{\epsilon}^T \boldsymbol{g}, \text{ where } \boldsymbol{g} = \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, y) \tag{3.3}$$

subject to $||\boldsymbol{\epsilon}||_\infty < \eta$. The solution to Equation 3.3 is:

$$\boldsymbol{\epsilon} = \eta \cdot \text{sign}(\boldsymbol{g}) \tag{3.4}$$

FGSM is extremely fast compared to the L-BFGS method of Szegedy et at. [273] since it required gradient computation only one time. The original version shown above creates an untargeted attack, since the user does not specify the required $\hat{y}$ of the adversarial example [102, 213]. FGSM can become targeted as follows:

$$\hat{\boldsymbol{x}} = \boldsymbol{x} - \eta \cdot \text{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \hat{y})), \text{ where } \hat{y} = \text{argmin}_y f_y(\boldsymbol{x}) \tag{3.5}$$

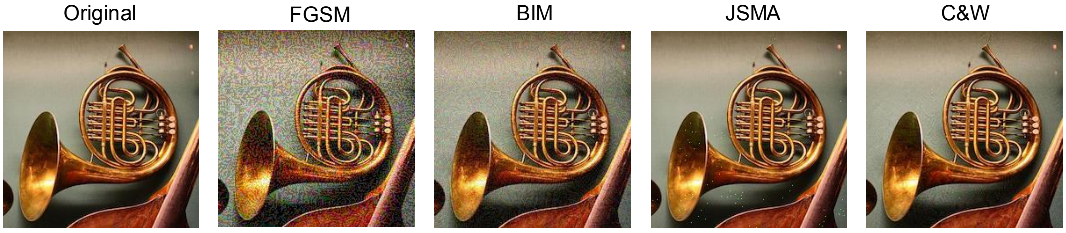


Fig. 3.3. Overview of adversarial examples generated with some of the discussed attack crafting methods, namely FGSM [102], BIM [164], JSMA [213] and C&W [51]. The image belongs to the ImageNet dataset [74] and for the generation *adversarial.js* [36] was used.

FGSM, after one step, usually required a high η to produce strong adversarial attacks; however, those can be detected by the human eye. Basic Iterative FGSM (BIM), where η is gradually increased, is another variation of FGSM that requires little computation with low overall perturbation [164].

Most methods of crafting adversarial examples perturb many input pixels, each by a minimal amount. Papernot et al. [213] introduced the Jacobian-based Saliency Map Attack (JSMA), a different approach that changes only a few input pixels, each one by a large amount. They extend the idea of saliency maps [255], commonly used as a visualization to compute adversarial saliency maps. These maps indicate which input pixels should be perturbed by the adversary to cause the desired changes to the network prediction. After each iteration, they maximally modify the highest-saliency pixels. JSMA is successful on input images with low dimension like MNIST [167] but is too computationally expensive for large images.

Madry et al. [183] proposed an extension of BIM called Projected Gradient Descent (PGD) attack and formulated it as a constrained optimization problem. The constraint for η is expressed as the L_2 or L_∞ norm of the perturbation. The difference in comparison to BIM is that PGD initializes the example to a random point in the ball of interest within the L_2 or L_∞ norm, while BIM initializes to the original point. PGD starts from a random perturbation in the ball around a sample. Then takes a gradient step in the direction that maximizes the loss, and it projects the perturbation back into the L_2 or L_∞ ball to maintain a minimal amount of added perturbation. The process is repeated until convergence [158, 183]. Carlini and Wagner introduce another strong constrained optimization-based attack (C&W) [51]. Examples of the aforementioned attacks can be seen on Fig. 3.3.

Another well-known adversarial attack is DeepFool [202], which consists of an iterative greedy search process. In every iteration, the projections of the input image to the decision boundaries of all classes are computed, and η is inferred to push $\boldsymbol{x}$ towards the decision boundary of the closest incorrect class. An extension of DeepFool is Universal Adversarial Perturbation [201]. The method computes a universal perturbation for a set of training samples by aggregating individual perturbation vectors that send the input samples towards the decision boundary. They showed that such perturbations were successful within the images of one dataset and were transferable among network architectures. In a similar direction, Brown et al. [41] introduced the Adversarial Patch, a universal, targeted adversarial attack. Adversarial patches can be printed and added to any scene or photograph, causing misclassified predictions to various classifiers. It should be noted that the patch, even though it is small, is not imperceptible and can be detected by human observers.

3.2.2 Transformation-based Attacks

Xiao et al. [298] introduced the Spatially Transformed adversarial examples. They maximized the model's cost function to predict the target misclassification class and minimized the spatial transformation (flow) of the pixels in their neighborhood. Using this method, instead of changing individual pixel values, they were shuffling pixels within a small neighborhood. This attack was particularly robust against a common defense strategy called adversarial training, encouraging the attention of the model to be located in the wrong parts of an image. Another transformation-based approach is ManiFool [143], which finds optimized affine transformations that, when applied to an image, can fool a classifier. This attack is particularly interesting for this dissertation, as it is among the few ones that don't rely on individual pixel perturbations. The contribution, which will be described in the next Chapter, extended the idea of ManiFool and showed that ManiFool adversarial training could dramatically increase the performance of a model to random affine and projective transformations and the clean test set [217]. Recently, Rahmati et al. [228] proposed a Geometric Decision-based Attack (GeoDA), a geometric framework that linearizes the decision boundary of classifiers.

3.2.3 Other Attack Mechanisms

Su et al. [267] introduced the One-pixel attack. They used an Evolutionary Algorithm called Differential Evolution and iteratively generated adversarial examples that minimized the confidence of a classifier. Initially, several adversarial candidates are generated by modifying a random pixel. After getting the model's predictions, the previous pixels' positions and colors were combined, generating more adversarial candidates. This step is repeated until an adversarial image with one pixel perturbed reduced the model's confidence, causing misclassification. This method showed that differential evolution is not computationally expensive and can be applied to various problems [286].

Another group of attacks is using Generative Adversarial Networks (GANs) to craft adversarial examples [297]. A generator, a discriminator, and an attacker are trained jointly [17]. The proposed attack has high success rates with low generation time. Cycle Consistency GANs have also been similarly employed to craft adversarial attacks [138]. It has also been shown that a CycleGAN could act as both an attack and a defense mechanism since it can generate clean samples from adversarial images.

3.2.4 Applications of Adversarial Attacks

Inspired by the attack crafting mechanisms described above, various applications of adversarial examples have been demonstrated in many fields and beyond images.

Attacks on semantic segmentation

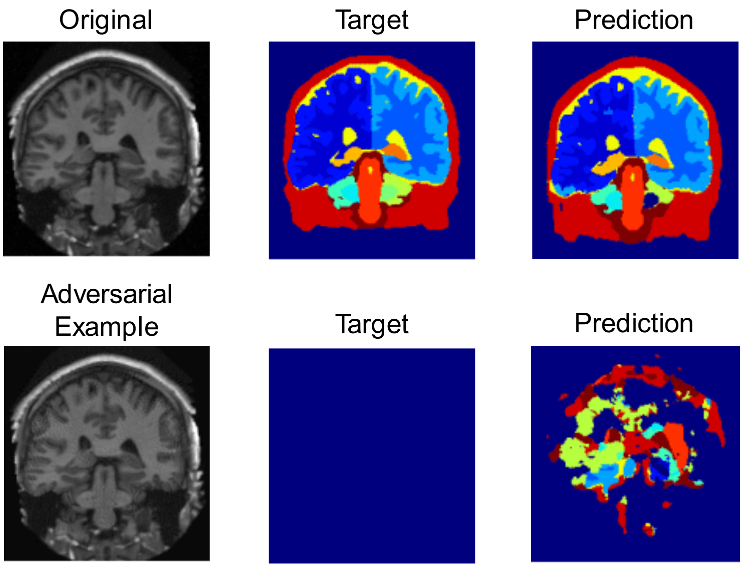


Fig. 3.4. Example of adversarial example crafted using DAG [301] in [215]. The target prediction is consists of solely background pixels and as can be seen by the model predictions, the attack is highly successful, substantially distorting the model prediction.

Xie et al. [301] introduced Dense Adversarial Generation (DAG), which operates like a per-pixel targeted version of FGSM. They showed that both semantic segmentation and detection networks could be fooled with high success rates. Especially for segmentation, regardless of the image's content, the predicted segmentation maps highly resembled the target ones that were completely different from the input image. In the next part of this dissertation, we will discuss one of our contributions [215] that leveraged DAG to generate adversarial examples for whole-brain segmentation that were then used as a benchmark for model robustness evaluation. An example of the DAG attack on whole-brain segmentation is shown in Fig. 3.4 using a slice from a volume from the OASIS [186] dataset. Fischer et al. [87] extended the universal adversarial perturbations for semantic segmentation and utilized them successfully, showing that a network can be tricked into not segmenting pedestrians standing in the middle of a street.

Attacks on audio and speech

Adversarial examples can be found in modalities other than imaging, specifically audio and speech. Alzantot et al. [7] proposed a method based on genetic algorithms to craft adversarial examples. The targeted attack creates a population of candidate adversarial samples by adding random noise to a subset of them within an audio clip. Afterwards, the fitness score for each population member is computed, and the next generation is produced, minimizing the noise effect on human perception.

Cisse et al. [66] introduced Houdini, a gradient-based approach that was successfully applied to speech recognition, pose estimation, and semantic segmentation in both a targeted and untargeted setting. Carlini and Wagner [48] introduced a white-box iterative optimization-based attack for speech recognition that measured distortion in Decibels (dB) and ensured it was imperceptible by humans with a very high success rate.

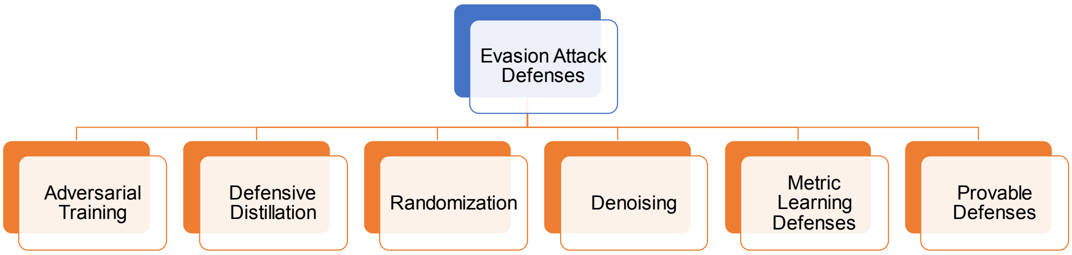


Fig. 3.5. Overview and taxonomy of defense mechanisms against evasion attacks with adversarial examples discussed in this dissertation.

Various methods have shown that Graph Convolutional Networks (GCNs) are also vulnerable to adversarial examples [31, 316]. Moreover, attacks have been proposed against deep reinforcement learning [24, 130, 229]. We won't delve into these approaches since they are out of the scope of this dissertation.

3.3 Adversarial Defenses

Adversarial robustness is useful for applications beyond security [209] and developing mechanisms to improve it can lead to substantially better models. Adversarial defenses vary significantly based on the attacks they are targeting, i.e. poisoning attacks during training time or test-time attacks. Furthermore, defenses can be categorized into **proactive**, which aim to create robust models before their deployment and **reactive** which aim to detect an attack on test-time and act accordingly. A taxonomy of defenses is shown in Fig. 3.5.

Adversarial Training

A widely used and straightforward approach against adversarial examples is to use them explicitly during training. This is a **proactive** defense strategy since models are protected against adversarial attacks during training before being deployed. It should be noted that adversarial training [102] is not equivalent to data augmentation [116] since adversarial examples are not images that are expected to occur naturally at test time and do not provide the same information to the network as the clean training images. Adversarial perturbations are recomputed using the latest version of the model parameters after every minibatch. The training process can be interpreted as a minimax game with the learning algorithm as the minimizing player and the adversarial crafting process (such as L-BFGS, FGSM or PGD) as the maximizing player [116]. Various defense methods incorporate adversarial training, either as the only measure [281] by augmenting training data with perturbations transferred from other models or as auxiliary support [240]. Moreover, experiments have been conducted on a large scale for adversarial training by Kurakin et al. [163].

The robustness provided by adversarial training relies heavily on the attacks used for training, which usually include FGSM and PGD. Since PGD is the most potent gradient-based attack, it brings the most significant improvement in robustness when used during adversarial training. A recent study by Athalye et al. [13] showed that adversarial training was one of the few defenses against adversarial examples that remained resilient.

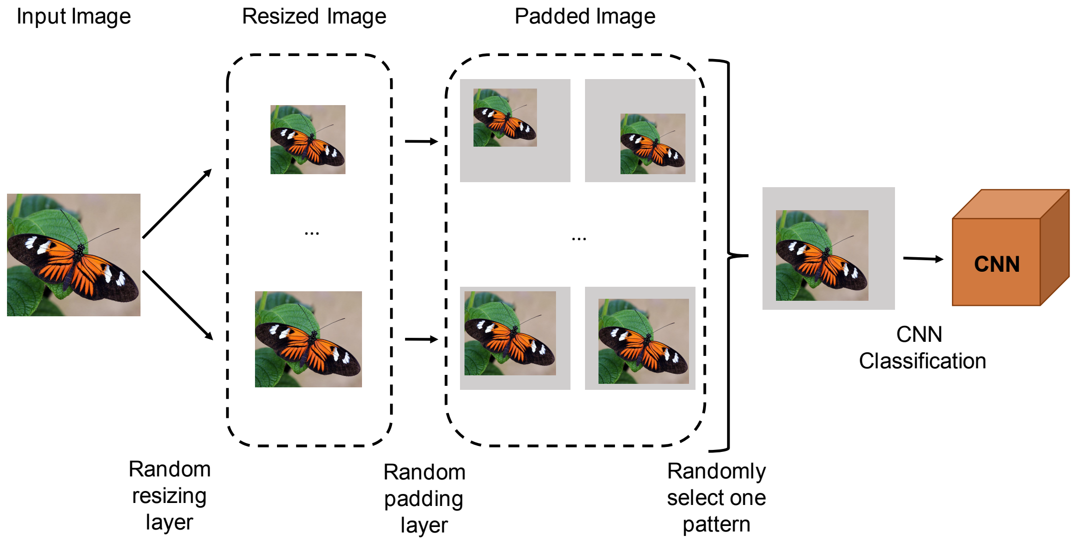


Fig. 3.6. Overview of the defense mechanism proposed in [300]. Random resizing and random padding are applied to an input image to eliminate the effects of adversarial perturbation before the model prediction.

However, adversarial training is time-consuming since, in addition to the gradients needed to update the model weights, each stochastic gradient descent iteration requires more gradient computations to craft the adversarial images. It has been shown that it takes 3-30 times longer to train a robust model with adversarial training than a non-robust equivalent. To that end, Shafahi et al. [248] introduced a fast adversarial training variation in which both network parameters and added image perturbations are computed once using a simultaneous backward pass. Later, adversarial training was also introduced for universal perturbations [249] as a min-max optimization problem where the minimization is over the model weights and the maximization over a universal perturbation.

Kannan et al. [145] proposed a different adversarial training technique called adversarial logit pairing (ALP). ALP encourages the similarity between pairs of images in the learned logit space by including the cross-entropy between the logits of benign samples and the corresponding perturbed samples in the training loss.

Defensive Distillation

Papernot et al. [211, 214] propose to combine two models in a student-teacher fashion to increase the robustness against adversarial examples inspired by knowledge distillation [125]. Specifically, the teacher network is trained with the original training dataset of clean images with one-hot ground truth labels and learns to predict a probability distribution over the labels of each class. Afterwards, the student model is also trained using clean training data, but instead of the ground truth for each image, the probability distribution predicted by the teacher model is used as a label. That way, the boundaries formed between the classes are less linear, and the model is encouraged to have lower confidence when classifying an ambiguous sample. However, this defense mechanism was found vulnerable by Calrini et al. [49]. Our experiments found that models trained with defensive distillation were vulnerable to black-box attacks but were substantially more robust to adversarial crafting, producing higher levels of perturbation for adversarial examples, creating more distorted and easily detectable samples.

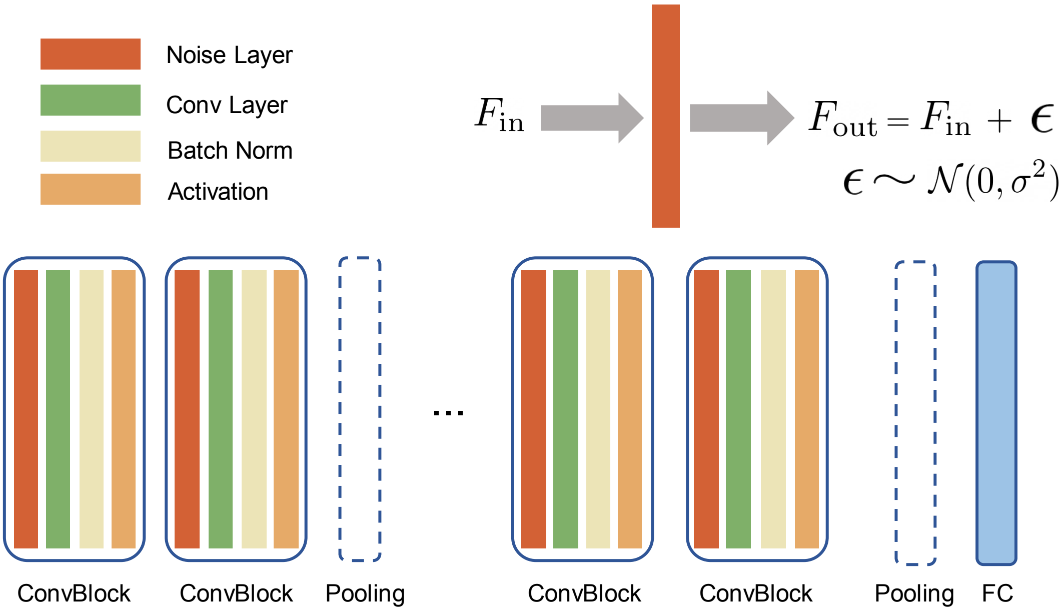


Fig. 3.7. Overview of the defense mechanism proposed in [178]. This method introduces a noise layer before each convolution during both train- and inference-time and ensembles the predictions for different random noise to stabilize the outputs of a model and eliminate the effect of adversarial perturbations.

Randomization

Recent defenses resort to randomization mechanisms to minimize the effects of adversarial perturbations in the input or feature space. Randomization-based defenses attempt to randomize the adversarial effects so that they are turned into random effects, which are not a concern for the majority of DNNs. Since adversarial perturbations result from meticulous optimization procedures, often dependent on a model and dataset, randomizing the input can eliminate the adversarial effects.

Xie et al. [300] utilize two random transformations, resizing and padding, to eliminate the adversarial effects at inference time as can be seen in Fig. 3.6. Random resizing resizes the input image to a random size before forwarding them to a trained model. Random padding added zeros around an input image in a random manner. This intuitive approach achieves robustness to black-box adversarial attacks. However, under the white-box setting, this mechanism was compromised by an attack proposed by Athalye et al. [14]. Guo et al. [107] apply random image transformations such as bit-depth reduction, JPEG compression, total variance minimization, and image quilting before feeding an image to a model. This defense approach is robust to black-box adversarial examples but is also vulnerable to the attack proposed in [14].

Liu et al. [178] introduce a random noising mechanism called random self-ensemble (RSE) shown in Fig. 3.7. RSE adds a noise layer before each convolution layer during both train- and inference-time and ensembles the prediction results over different random noises to stabilize the outputs of a model and minimize the effect of adversarial perturbations. Moreover [172] proposes to directly add random noise to image pixels before classification in order to eliminate

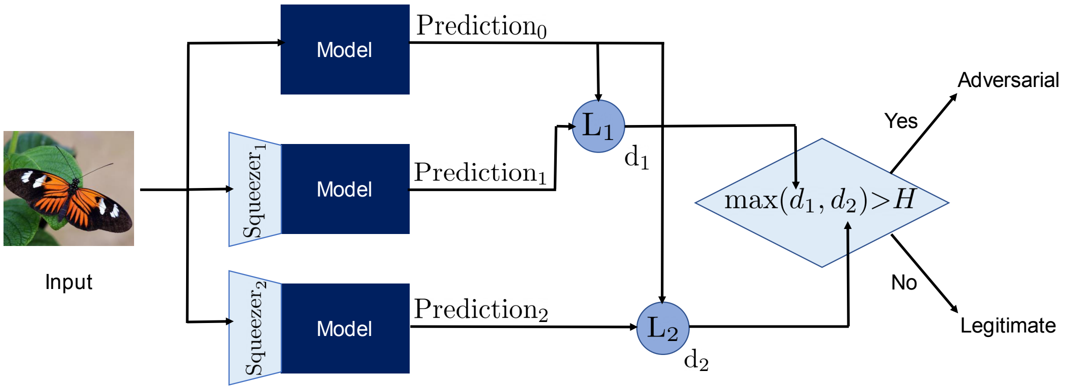


Fig. 3.8. Overview of the adversarial example detection method proposed in [304]. The network is evaluated on the original input and input pre-processed by the feature squeezers. If the difference between the model's prediction on a squeezed input and its prediction on the original input is higher than a threshold, the input is considered adversarial.

the strength of adversarial perturbations and provides an upper bound regarding the size of adversarial perturbation the method is robust to.

Dhillon et al. [77] introduce stochastic activation pruning (SAP) to defend pre-trained models against adversarial examples by stochastically pruning a subset of the activations in each layer while maintaining the activations with larger magnitudes. Even though this method is robust to some black-box attacks, it is also vulnerable to the attack of Athalye et al. [14] that is tailored to randomization defenses. Luo et al. [181] propose to mask the feature maps after each convolutional layer randomly. By randomly masking these maps, each filter only extracts features from partial positions. This defense shows high robustness to black-box attacks, and the adversarial examples it is vulnerable to are usually also confusing to humans.

Denoising

Denoising is a straightforward **reactive** method to mitigate the effects of adversarial perturbations. Previous works can be categorized into input denoising and feature denoising. Input denoising attempts to remove the adversarial perturbations from an input image directly, while feature denoising alleviates the effects of adversarial perturbations directly on DNN features.

Xu et al. [303, 304] introduce two denoising methods using bit-reduction and image-blurring to reduce the effects of adversarial perturbations. Adversarial example detection is achieved by comparing the predictions of a model on an original and a squeezed image as can be seen in Fig. 3.8. The original input is considered an adversarial example if the two types of inputs produce substantially different predictions. However, feature squeezing was shown to be vulnerable to adversarial attacks of increasing difficulty [118, 250].

Other works use GANs to learn the distribution of the clean data to generate a clean projection of an adversarial example, thus alleviating the perturbation effect. Defense-GAN [241] trains a generator to model the distribution of clean images. During inference, Defense-GAN denoises an adversarial example by looking for an image close to the adversarial input in its learned

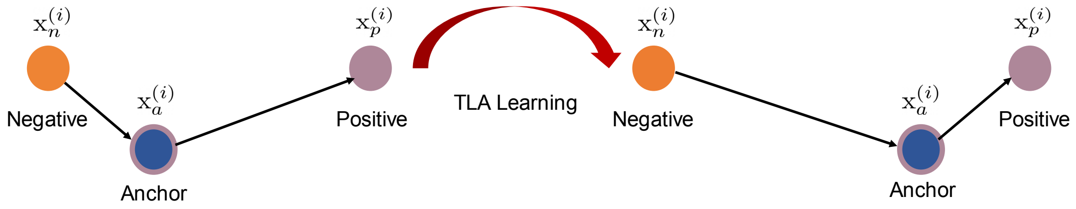


Fig. 3.9. Overview of the metric learnign defense approach proposed in [185]. In Triplet Loss Adversarial (TLA) training a triplet is selected so that the negative sample is the closest one withing the mini-batch. After training the distance between clean and adversarial examples originating from the same class is minimized while the margins to different classes increase.

distribution and forward this benign image into the model instead. This method remains highly effective against adversarial attacks, and its success has only been reduced by [13].

Similarly, the adversarial perturbation elimination GAN (APE-GAN) [139] also trains a generator to denoise an adversarial example. Even though APE-GAN achieves a good performance in a black-box setting, it is vulnerable to the adaptive white-box attack introduced in [50].

In MagNet [195] an auto-encoder is leveraged to learn the manifold of clean samples. Afterwards, a detector distinguishes between clean and adversarial examples according to the relationships between those and the learned manifold. The reformer is trained to denoise the adversarial samples and turn them to their clean counterparts. Despite MagNet's success against black-box attacks like FGSM, Carlini and Wagner [50] showed that MagNet is vulnerable to transferable adversarial examples crafted with the CW_2 attack.

Liao et al. [175] describe a High-level representation Guided Genoiser (HGD) that denoises the features distorted by the adversarial perturbations. Instead of denoising the image pixels, HGD uses a denoising U-net [236] with a feature-level loss function that minimizes the feature difference between clean and adversarial examples. Even though HGD was successful against black-box attacks, it was compromised by a PGD attack in a white-box setting [12].

Metric Learning Defenses

Mao et al. [185] introduced Triplet Loss Adversarial (TLA) training, a metric learning defense. For a triplet, the negative example was selected to be the image closest to the anchor that belongs to a different class as shown in Fig. 3.9. In addition, a positive sample from the same class as the anchor was randomly selected within a mini-batch. Thus, TLA brought clean and adversarial examples originating from the same class closer and increased the margins to the different classes. TLA showed increased robustness even for attacks such as PGD.

Papernot et al. [210] proposed a defense based on k-nearest neighbors (KNN) called Deep KNN (DkNN). KNN is run on the representations of each layer of a DNN and is used to estimate the abnormality of a prediction for a test sample. The prediction can be considered abnormal when the DNN representations for a test input are far from the representations of training samples of the same class as the predicted one. DkNN showed its robustness to various adversarial attacks, including the C&W attack.

Provable Defenses

The defenses described above are based on heuristics, which means that their effectiveness is experimentally validated, rather than theoretically proved. Thus, even though some of those heuristic defenses are still robust, they could still be vulnerable to future attacks. To that end, we will now discuss provable defense mechanisms that maintain their robustness for a well-defined type of attack.

Raghunathan et al. [226] propose a certifiable defense method against adversarial examples for two-layer models. The authors propose a semidefinite relaxation that can create a differentiable certificate for the network's robustness. Afterwards they incorporate the relaxation to the loss as a regularizer to encourage robustness. This method certifies that no attack that perturbs each image pixel by at most $\epsilon = 0.1$ can cause more than 35% test error on MNIST [167]. They later expanded their certificate to more networks, also including the ReLU activation function [227].

Wong and Kolter [294] formulate a dual problem to upper-bound the adversarial polytope. Different from [226] this approach can be applied to deep networks with arbitrary linear operators, such as convolutions. Their approach is scaled further in [295] to more general deep architectures with skip connections and nonlinear activations. They also propose a nonlinear random projection technique to estimate the upper bound that only scales linearly with the size of hidden units, making it applicable to larger models.

Balunovic et al. [21] proposed Convex Layerwise Adversarial Training (COLT) which combines adversarial training and provable defenses. Specifically, the training procedure combines both the verifier and the adversary. The verifier aims to certify the model with convex relaxation while the adversary attempts to find inputs inside the convex relaxation, which cause verification failure.

Another approach to provable defense is introduced in [258], where the problem is tackled as distributionally robust optimization. They consider a Lagrangian penalty formulation of perturbing the underlying data distribution in a Wasserstein ball and follow a training procedure that combines model weight updates and worst-case perturbations for the training data.

3.4 Adversarial Examples in Medical Imaging

Adversarial robustness has also been an active area of research in medical imaging due to the need for increased security in critical healthcare systems. To this end, one of the first introductions of adversarial examples for medical imaging is one of the contributions of this dissertation [215]. In our work, which will be thoroughly discussed in the next Part, we proposed to use adversarial examples as benchmark for thorough model evaluation, especially in cases where different model architectures achieve the same performance on clean data. Finlayson et al. [86] discuss how the healthcare system may be uniquely vulnerable to adversarial attacks, both regarding monetary incentives and technical issues and attack state-of-the-art DNNs for the tasks in dermatology, ophthalmology and radiology.

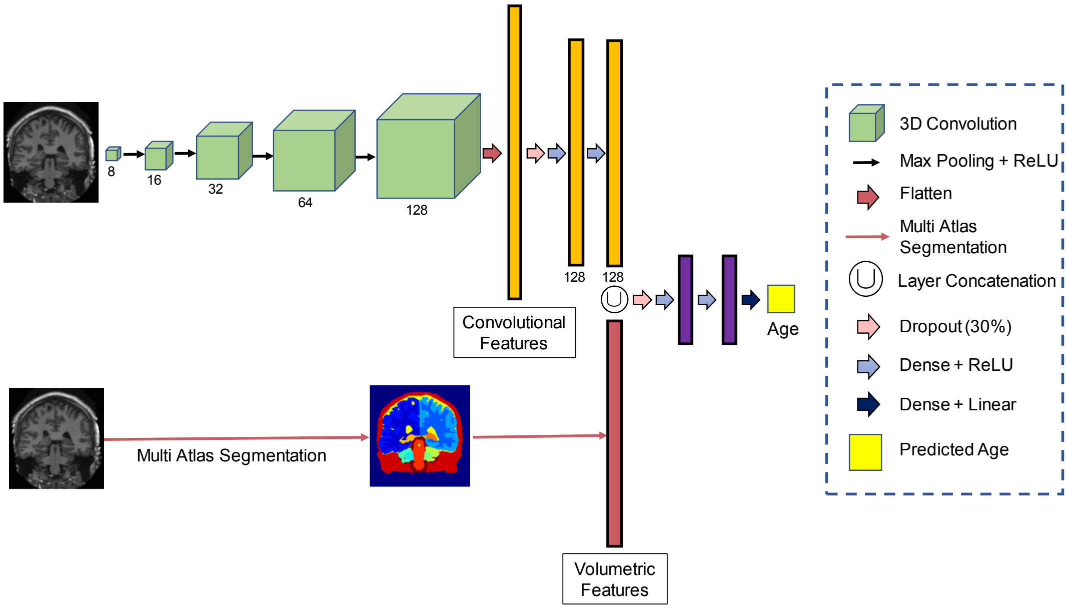


Fig. 3.10. Overview of the method proposed in [174]. By incorporating additional anatomical information, specifically the volume of each brain structure, the model's age predictions became more robust.

Notably, Ma et al. [182] found that medical DNNs can be more vulnerable to adversarial attacks compared to models for natural images. This is attributed to the fact that medical images usually have complex biological textures with high gradient regions that are sensitive to adversarial perturbations; and that state-of-the-art DNNs designed for large-scale natural image problems are often overparameterized, leading to a sharp loss landscape and high susceptibility to adversarial attacks. However they also found that, medical adversarial attacks can be easily detected, due to critical feature differences compared to natural images.

Adversarial Attacks for Medical Applications

Li et al. [306] proposed Adaptive Targeted Iterative FGSM (ATI-FGSM), a tailored attack against DNNs for multiple landmark detection. ATI-FGSM adds imperceptible perturbations to an image and can influence the model's predictions of user selected landmarks, while keeping the other landmarks still. Experiments showed that ATI-FGSM was more effective against the original Iterative FGSM attack for cephalometric landmark detection.

Commonly used attack mechanisms do not directly extend to Electrocardiogram (ECG) signals, since such methods introduce artefacts to the signals that are not physiologically possible. To that end, Han et al. [111] recently developed a method to construct smoothed adversarial examples for ECG tracings that are imperceptible to human evaluation. Their attack is evaluated on a DNN for arrhythmia detection from single-lead ECG, showcasing the vulnerability of the model to the newly introduced attack.

Adversarial Defenses for Medical Applications

Li et al. [174] proposed a hybrid model for age prediction from brain MRI scans that is shown in Fig. 3.10. Specifically, they found that introducing anatomical context to the training process

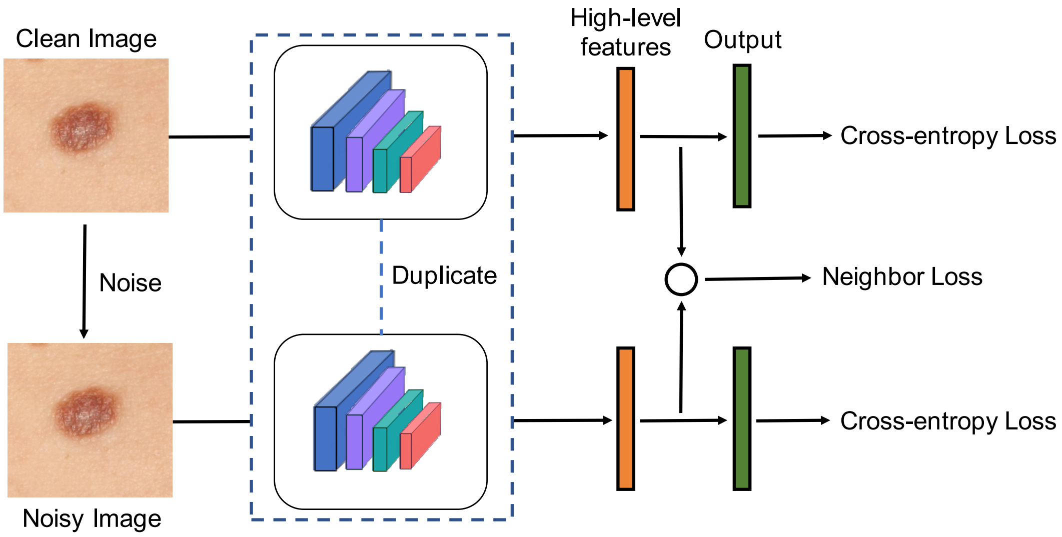


Fig. 3.11. Overview of the denoising method proposed in [305]. A neighbor loss is used to amplify the similarity between noisy and clean samples in the feature space and enhance the robustness of the model by denoising adversarial samples.

substantially increases the model's robustness. To this end, they augmented the training of the DNN with the volume of each brain region computed through traditional multi-atlas segmentation techniques. Another approach is utilizing adversarial training with PGD samples for lung nodule detection [177]. They showed that their method was more robust to both under-represented nodules and resilient to noise perturbations.

An adversarial example detection method was proposed in [173] for Chest X-ray multi-label pathology classification. For a clean or adversarial input, the system first extracts features using a CNN classifier trained on clean data. Afterwards, the detection module rejects the input if it is considered adversarial or predicts the classification label. The detection module uses a unimodal multivariate Gaussian model (MGM) to identify the attacks.

Bortsova et al. [292] analyzed black-box adversarial attacks for DNNs used for ophthalmology, radiology, and pathology. They investigate the effect of pre-training and training data similarity to a model's robustness. Their experiments showed that pre-trained weights could lead to higher adversarial example transferability and that data disparity between the target and the source models of the attacks contributed to increased robustness. Their analysis concluded with a set of recommendations to increase the robustness of a system deployed for healthcare.

Another defense strategy, shown in Fig. 3.11 was described in [305], where they developed a method that directly enhances a classifier's denoising ability with a naturally embedded auto-encoder and a mechanism for semantic feature invariance for general noise. Specifically, a neighbor loss is employed to emphasize the similarity between noisy and clean samples in the feature space. Experiments on dermatology and radiology showed that the robustness of a classifier with the proposed denoising strategy could be substantially improved. Xu et al. [302] introduced two defense strategies based on adversarial training, namely, Multi-Perturbations Adversarial Training (MPAdvT) and Misclassification-Aware Adversarial Training (MAAdvT). MPAdvT trains DNNs using different perturbation levels and varying adversarial iteration steps.

In MAAdvT a misclassification-aware regularization using Kullback-Leibler (KL) divergence is added to the adversarial loss to stabilize the loss for clean samples that are being misclassified by the DNN.

Explainability with Adversarial Examples

Recent works aim at using adversarial examples to improve various aspects of model training, evaluation and interpretation. Khakzar et al. [150] aim at enhancing the learned feature representations of DNNs by training models that are robust against adversarial examples. Their approach using robust optimization steers the model towards learning more interpretable features. They evaluate their method on weakly-supervised localization of anomalies on Chest X-Rays showing increased localization accuracy and enhanced interpretable gradients.

Chang et al. [53] proposed an adversarial explanation technique for applications in ophthalmology. Their regularization method, inspired by the Lipschitz constraint, showed that when the model is distorting images to deliberately change the prediction to pathologic or normal it is providing explanations for the DNN decision process. Glaucoma specialists compared conventional heatmap-based explanation methods with the proposed adversarial explanation and identified substantial improvement in the explainability of the model.

Geometric Transformations

4

In this Chapter we are discussing the first contribution of this dissertation which has been published in [217]. Figures 4.1-4.4 and Tables 4.1-4.4 are used with permission from Springer Nature Customer Service Centre GmbH with License Number: 5058290266725.

4.1 Introduction

Deep learning models have recently been effective in performing medical imaging tasks such as classification, segmentation, and registration with state-of-the-art accuracy and have found their way into a multitude of Computer Assisted Diagnosis and Intervention (CAD/I) Systems that assist physicians.

However, medical imaging datasets are often characterized by considerable class heterogeneity, extreme class imbalance, outliers, inter-observer variability, ambiguity, and, most importantly, limited data. The aforementioned issues hinder neural network training, resulting in suboptimal and overfit solutions.

Furthermore, deep learning models used by physicians in a CAD/I system must be extensively evaluated, not only in terms of generalizability (performance on data originating from a given test set), but also in terms of robustness (behavior on data corrupted by noise, unknown transformations, and outliers). Data augmentation is the process of increasing the size and variance of a dataset used to train a machine learning model aiming to improve generalizability and gain a deeper understanding of the underlying distribution of the training data. The space representing the distribution of the training data can be viewed as the manifold of a given class learned by a classifier.

ManiFool Augmentation is achieved by increasing the size of the training dataset for a given task with samples transformed with optimized affine geometric transformations. The overview of the method is shown in Fig. 4.1, where it is compared with traditional data

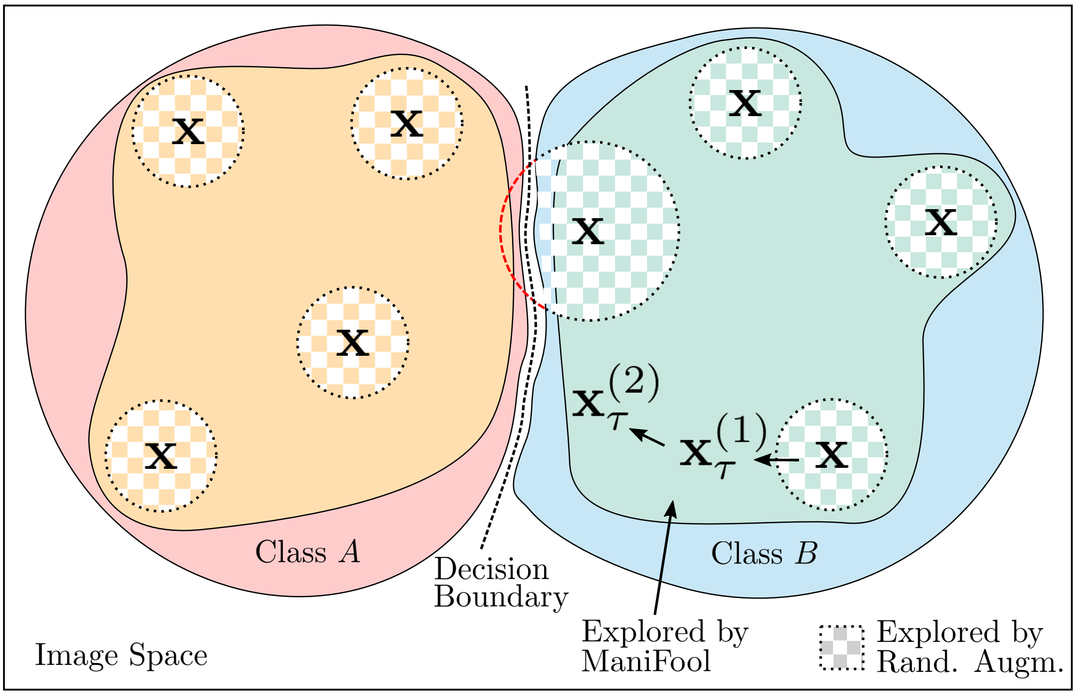


Fig. 4.1. **Schematic representation of proposed augmentation**: In contrast to random augmentation (checkerboard pattern), which explores the space around the original training samples $\mathbf{x}$ locally, the proposed augmentation scheme based on ManiFool explores the present classes towards the decision boundaries, thus incorporating more relevant training samples $\mathbf{x}_{\mathcal{T}}^{(i)}$. Furthermore, ManiFool Augmentation samples are ensured to belong to the ground truth class. Figure published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

augmentation performed with random transformations. The algorithm used to create samples for data augmentation is inspired by ManiFool [143], and the intuition behind it is rather intuitive: Iteratively move an image towards a classifier's decision boundary using affine geometric transformations, following the direction that maximizes the gradient. After every step, project the computed movement back onto the original training manifold of the image being transformed. This procedure is repeated until either a transformation is found that causes the network to misclassify the transformed sample or a maximum allowed number of steps is reached. In case of misclassification, we have crossed the decision boundary and entered the manifold of another class. In that case, we backtrack to the manifold of the original class and leverage the computed affine transformation for data augmentation.

Unlike conventional augmentation approaches that use random transformations, ManiFool Augmentation ensures that the network's training space is not restricted to the immediate vicinity of a training sample. Instead, as shown in Fig. 4.1, augmentations are found globally up to the limits of each class-manifold for the entire training set. An effective data augmentation technique should ensure that the samples used to increase the population of the training dataset originate from the same manifold as the original data. Using training samples from a different distribution in the training dataset for augmentation would not necessarily allow the model to learn a better embedding for each class but rather map the same class to two different sub-spaces, one for each training manifold.

A thorough evaluation on two challenging medical datasets shows that the proposed augmentation method improves the robustness of a model to geometric transformations and substantially increases the performance on the original test data. This is further underlined by cross-dataset testing, where networks trained with ManiFool Augmentation better captured the underlying distribution of the training data.

Contributions:

- We propose a novel **data augmentation technique**, utilizing an exhaustive manifold-exploration method that improves the performance of a deep learning model on the provided test set and substantially increases its robustness to random geometric transformations.
- We describe **quantitative measures** to evaluate the robustness of a classifier. A metric like this is a step toward rigorous evaluation of machine learning models, a significant move toward physicians' safe and accurate use of trained models in practical applications involving patient diagnosis and care.
- We compare three state-of-the-art DNN architectures in terms of their robustness to geometric transformations and evaluate the quality of their decision boundaries.
- We thoroughly validate the proposed augmentation technique and robustness metric on two challenging medical imaging datasets for skin and breast lesion classification.

4.2 Related Work

General Applications

Supervised learning relies heavily on the training dataset used to optimize a classifier. Specifically, in medical imaging applications, datasets can be small, with noisy annotations and bias to a specific site, scanner, or population group [147]. Thus, utilizing general methods to reduce the model memorization of the training dataset and avoid overfitting is critical. As will be discussed later in this dissertation, models that better understand the underlying manifold of the training data showcase increased robustness.

A widely used method to avoid overfitting and enhance a model's performance is data augmentation with random transformations [254]. These transformations can include scaling, rotation, translation, and flipping. Furthermore, adding salt and pepper or Gaussian noise can simulate images of decreased quality. Perspective transforms are also used that project the image from different points of view to emphasize different objects of significance.

Methods range from elastic transformations [296], noise generation in a learned features space [76], to repeat, rotate and infill approaches where scaling and rotation are applied to a training sample in a grid pattern, and background consistency is guaranteed [208]. Fawzi et al. introduced a data augmentation method that can be used with stochastic gradient descent and looks for an augmented sample with the greatest loss within a constrained exploration space called "trust region" [84].

Generative Adversarial Networks (GANs) [103] are also widely used to produce synthetic data augmentation samples for various applications. Data Augmentation GAN (DAGAN) [10] learns how to craft synthetic images using a lower-dimensional representation of real examples. In DAGAN, the generator acts as an autoencoder; it encodes a given image to a compact representation, adds noise to it, and then decodes it. Thus, the decoder learns a family of transformations that can act as data augmentation. The DAGAN discriminator learns to distinguish between an image and a transformed version of it and a pair of different images from the same class. That way, the discriminator encourages the decoder to learn transformations that do not alter the class but produce transformed images that differ substantially from the original image.

BAlancing GAN (BAGAN) [188] is a conditional GAN approach tailored towards datasets with severe class imbalance. The model learns features from the majority classes and uses them to generate images for the underrepresented classes. Unlike DAGAN, class conditioning is applied to the latent space to push the generation towards a target class. Wang et al. [289] reported, however, that in many cases, traditional data augmentation with geometric transformations outperforms GAN-based approaches and is more generalizable.

Other methods for data augmentation that additionally increase robustness include Patch Gaussian [179], where Gaussian noise is added to randomly selected patches of an image. This method can increase robustness to high-frequency noise while maintaining the ability to take advantage of high-frequency information to classify an image correctly. Such a general method can also be used in combination with other techniques such as AutoAugment [70]. AutoAugment poses augmentation as a discrete search problem in which the search algorithm is based on reinforcement learning, which aims to maximize the classification accuracy with data augmentation.

To improve robustness for in-domain and Out-of-Distribution (OOD) data, Self-Supervised Manifold Based Data Augmentation (SSMBA) has been proposed [206], generating synthetic training samples by a pair of corruption and reconstruction functions to move randomly on a data manifold.

Medical Imaging

For medical imaging applications, all data augmentations applied to a dataset must be anatomically and physically correct. Populating the training dataset with highly distorted or infeasible images can lead to model ambiguity and hinder model convergence and generalizability.

Nalepa et al. [205] proposed a data augmentation method that exploits diffeomorphic image registration to benefit from subtle spatial and tissue features captured within the training set for brain-tumor segmentation from MRI scans. A method for one-shot biomedical image segmentation with learned image transformations was introduced in [310] where a single segmented scan is required and combined with other unlabeled scans in a semi-supervised approach learns a model of transformations and synthesizes additional labeled samples. The transformations are composed of a spatial deformation field and an intensity change, creating variations in anatomy and image acquisition procedures. Chen et al. [54] introduced a data augmentation technique to learn both generalizable and robust features for cardiac MR image segmentation using an adversarial intensity transformation model, which can simulate intensity inhomogeneities, a common artefact in clinical MR imaging.

Tirindelli et al. [280] introduced ultrasound-tailored data augmentation transformations inspired by the physics of the modality. Specifically, they proposed deformation, reverb and Signal-to-Noise Ratio transformations that were applied to B-Mode ultrasound images of the spine. The method was validated for segmentation and classification tasks and showed promising improvement in comparison to DNNs trained with no or with random augmentation.

GANs are also widely used in the field of medical imaging for a variety of augmentation techniques. Bowles et al. [33] generated high-quality data augmentation samples along with their annotations using Progressive Growing GANs (PGGANs) for CT and MRI scans. In [92] Deep Convolutional GANs (DCGANs) were used to increase performance on liver segmentation from CT scans with synthetic images. GANsfer Learning [34] combined labeled and unlabelled data with a PGGAN and decoupled the learning of structural variations and the learning of structure appearance. This method produced high-quality augmentation images in 3 learning phases used for grey matter segmentation from MRI scans. Another commonly used framework was based on Cycle Consistency GANs (CycleGANS) [313] to transform contrast CT images into non-contrast images [242]. The synthetic non-contrast images were leveraged as data augmentation for kidney segmentation from CT scans, improving the model's performance and showcasing increased robustness to OOD samples.

Contrary to the approaches described above, ManiFool does not utilize GANs to create synthetic samples but uses an exhaustive manifold-exploration method to find the optimal affine transformations that can increase a model's performance and robustness to geometric transformations.

4.3 Methodology

ManiFool [143] is an iterative algorithm that can be applied to any differentiable classifier f. This section will describe the mathematical operations that generate a geometrically transformed example leveraged for data augmentation.

Movement Direction

With an image $\mathbf{x}$ with ground truth label y and a binary classifier f we initialize an iterative process of i steps. The input image at step 0 is noted as $\mathbf{x}^{(0)}$. First, ManiFool calculates the movement direction $\mathbf{u}$ to the decision boundary of classifier f. This is done by following the opposite of the gradient, $-\nabla f(\mathbf{x})$. The gradient at the step i for the image $\mathbf{x}^{(\mathbf{i})}$ is the projection of $\nabla f(\mathbf{x}^{(\mathbf{i})})$ onto the tangent space and can be calculated with the pseudoinverse given by:

$$\mathbf{u} = -\mathbf{J}^{+}_{\mathbf{x}^{(\mathrm{i})}} \nabla f(\mathbf{x}^{(i)}) = -(\mathbf{J}^{\mathbf{T}}_{\mathbf{x}^{(\mathrm{i})}} \mathbf{J}_{\mathbf{x}^{(\mathrm{i})}})^{-1} \mathbf{J}^{\mathbf{T}}_{\mathbf{x}^{(\mathrm{i})}} \nabla f(\mathbf{x}^{(i)}). \tag{4.1}$$

$\mathbf{J}_{\mathbf{x}^{(\mathrm{i})}}$ denotes the Jacobian matrix and the computed $\mathbf{u}$ for step i defines the direction towards the decision boundary.

To increase the accuracy and speed-up the convergence during the calculation of $\mathbf{u}$ a manifold optimization method following [3] was used:

$$\mathbf{u}^{(i)} = -\lambda_i \frac{\mathbf{J}^{+}_{\mathbf{x}^{(\mathrm{i})}} \nabla f(\mathbf{x}^{(i)})}{||\mathbf{J}^{+}_{\mathbf{x}^{(\mathrm{i})}} \nabla f(\mathbf{x}^{(i)})||} + \gamma \mathbf{u}^{(i-1)}, \tag{4.2}$$

where λ_i denotes the computed step size in the current iteration and γ is a constant momentum.

Mapping onto the original manifold

Once the movement direction $\mathbf{u}$ towards the decision boundary is computed, it is mapped back onto the manifold of the ground truth class denoted by $\mathcal{M}$. As in [143], the mapping is carried out with retraction $R_{\mathbf{x}^{(\mathrm{i})}}(\mathbf{u}) = \mathbf{x}^{(\mathbf{i})}_{\tau_{\mathrm{i}}}$, where τ_i is the affine transformation computed:

$$\tau_i = \exp\left(\sum_j u_j Gj\right). \tag{4.3}$$

G_j denote the basis vectors of the Lie Group $\mathcal{T}$ of the computed affine transformation. The algorithm terminates once one of the following two conditions are met: a transformed image has been misclassified, or the maximum number of iterations $I_{\max}$ was reached. After i steps the transformation applied to the input image $\mathbf{x}^{(\mathbf{0})}$ to generate the ManiFool sample are accumulated and calculated as:

$$\hat{\tau} = \tau_0 \circ \tau_1 \circ \ldots \tau_i. \tag{4.4}$$

Multi-class Classification

In order to extend ManiFool to multi-class classifiers, we execute the following steps: We craft a ManiFool example for each of the remaining classes. We calculate the geodesic distance between the original and the transformed image for each class. Afterwards, we select the generated example, which requires the smallest transformation $\tau_{y_{\min}}$ to cause an erroneous prediction. The class with the lowest geodesic distance between the original and transformed image can be calculated as:

$$l_{\min} = \arg\min_{l \neq l_x} \tilde{d}_{\mathbf{x}^{(0)}}(e, \tau_l). \tag{4.5}$$

Next, we will elaborate as to how the distance $\tilde{d}_{\mathbf{x}^{(0)}}$ is computed and how it can be used as a metric for robustness evaluation of DNNs.

4.3.1 Invariance to Geometric Transformations

Geodesic Distance Between Transformations

The geodesic distance $d_{\mathbf{x}^{(i)}}$ between two transformations τ_1 and τ_2 is given by the length L of the shortest curve γ between τ_1 and τ_2. However, the metric space of the manifold of the training data is not known. Thus, we have to acquire a metric in the Riemannian space. This can be achieved by mapping the Lie group $\mathcal{T}$ to the differentiable image manifold of $\mathbf{x}^{(i)}_{\tau_1}$ and $\mathbf{x}^{(i)}_{\tau_2}$, which inherits the Riemannian metric from L_2 [143, 160, 285]. The geodesic distance computed after the mapping between τ_1 and τ_2 is found by the shortest path between $\mathbf{x}^{(i)}_{\tau_1}$ and $\mathbf{x}^{(i)}_{\tau_2}$. This can be computed as:

$$d_{\mathbf{x}^{(i)}}(\tau_1, \tau_2) = \min L(\gamma). \tag{4.6}$$

Geodesic Distance Between Original Images and ManiFool Examples

After discussing how to compute the distance between two transformations and two transformed images, we will now elaborate on how to calculate the geodesic distance between the original images of the training set and the samples generated with ManiFool. The sample $\mathbf{x}^{(0)}$ before being transformed can be considered the initial point of the γ curve we mentioned above, if we define its transformation e as identity [143]. Hence, the geodesic distance between the original image $\mathbf{x}^{(0)}_e$ and $\mathbf{x}^{(i)}_{\tau_i}$, can be computed by the distance between the identity transformation e and the final aggregated affine tranformation τ_i as shown below:

$$\tilde{d}_{\mathbf{x}^{(i)}}(e, \tau_i) = \frac{d_{\mathbf{x}^{(i)}}(e, \tau)}{||\mathbf{x}^{(i)}||_{L^2}}. \tag{4.7}$$

Robustness to Geometric Transformations

The computed ManiFool examples are crafted so that they will originate from the edge of each class manifold. Thus, if we measure this distance $\tilde{d}_{\mathbf{x}^{(i)}}$ between an original input image and its ManiFool transformed example, we can define a metric to evaluate the robustness of a model. Specifically, we hypothesize that models with a learned feature space with increased class compactness and maximized distance among decision boundaries will require a higher average $\tilde{d}$ to cause a transformed image to be misclassified. We calculate the average distance $\tilde{\rho}_\tau$ of all the crafted ManiFool examples as:

$$\tilde{\rho}_\tau(f) = \frac{1}{m}\sum_{j=1}^{m} \tilde{d}_{\mathbf{x}^{(i)}_j}(e, \tilde{\tau}), \tag{4.8}$$

where m denotes the number of all crafted examples. $\tilde{\rho}_\tau$ can be used as a quantitative metric of the robustness of a classifier to geometric transformations. This metric can be leveraged to compare the robustness of different model architectures.

An additional metric for the robustness quantification of classifier f is r_τ, which can be computed by Equation 4.9. r_τ evaluates the performance of a model on images transformed randomly. In this work, for a range of given geodesic distances r we craft examples using random transformations and evaluate the misclassification rate of the model f.

$$r_\tau(f) = \min r \text{ s.t. } \mathbb{P}(f(\mathbf{x}^{(\mathbf{i})}_\tau) \neq f(\mathbf{x}^{(\mathbf{i})}) \mid d_{\mathbf{x}^{(\mathbf{i})}_\tau}(e, \tau) = r) \geq 0.5, \tag{4.9}$$

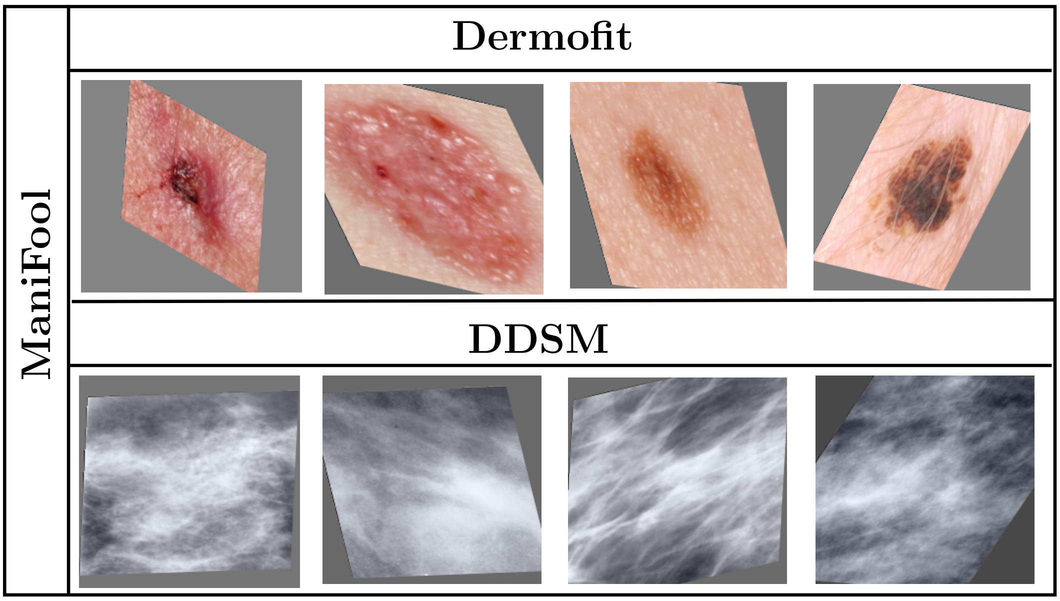


Fig. 4.2. Samples generated with ManiFool Augmentation for the two datasets, specifically Dermofit and DDSM. Figure published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

where 0.5 is a threshold defined by the user. It is hypothesized that a robust model can sustain higher classification accuracy for samples with higher geodesic distance from the original images, thus being more resilient to transformations.

4.3.2 ManiFool Augmentation

Our approach differs from the original ManiFool work in that our goal is to use the transformed images for data augmentation rather than fooling a deep neural network and crafting an adversarial example [273]. Thus, after we compute the affine transformation τ_i that crosses the decision boundary and causes misclassification for f, we backtrack onto the original class manifold $\mathcal{M}$ by iteratively reducing the final step size.

For the images in the training set, we craft ManiFool Augmentation examples that originate from the edges of the class manifolds. For this task, we utilize an independent black-box classifier f. Afterwards, we create an augmented training dataset consisting of original and geometrically transformed samples in an equal ratio and train a new randomly initialized classifier. An alternative method would have been to leverage every geometrically transformed sample generated at each step i for data augmentation. However, it was important to keep an equal ratio of transformed and original samples in the final dataset to avoid bias to geometrically transformed samples. Therefore, we only employed the transformed samples close to the decision boundary to accommodate the maximum possible variance during training without inducing bias to the classifier. Samples created with ManiFool Augmentation are shown in Fig. 4.2.

4.4 Experiments

4.4.1 Experimental Setup

Dataset Description

Two challenging medical imaging classification datasets were used to evaluate the proposed method; namely, Digital Database for Screening Mammography (DDSM) [119], [120] and Dermofit [18]. The DDSM database contains 11.617 expert-selected regions of interest (ROI) from mammograms from 1861 patients who were annotated by radiologists as normal, benign, or malignant. Dermofit is an image collection with 1300 high-quality dermatoscopic images and fine-grained expert annotations that have been histologically verified (10 classes). Both datasets were split at patient-level with non-overlapping folds; specifically, 70% was used for training and 30% for testing.

Model Training

Three state-of-the-art model architectures, namely ResNet18 [117], VGG16 [256] and InceptionV3 [271], were selected for the evaluation. All networks were initialized with ImageNet weights; therefore, appropriate resizing and normalization of the input were performed. The loss function used for the classification problems was weighted cross-entropy since the aforementioned datasets are characterized by severe class imbalance. Median frequency balancing was used to compute the class weights, following [238]. All the models were optimized with Adam optimizer with an initial learning rate of 0.001. The experiments were implemented in PyTorch [218], and the models were trained on an NVIDIA Titan Xp for 50 epochs.

Baseline Methods

We use ablative experiments as well as comparisons to other commonly used augmentation methods to evaluate the proposed contributions. The proposed method was compared with models trained with no data augmentation (referred to as "None" in the following Section) and models trained with traditional random augmentation ("Random"), specifically rotation and horizontal flipping. ManiFool Augmentation (noted as "ManiFool" in the tables of results) was further evaluated against augmentation techniques including Random Erasing [311] ("Erasing"), a commonly used and fast augmentation technique that replaces random patches of an image with Gaussian noise and data augmentation with images synthesized by GANs ("DCGAN"), following the method proposed in [92].

ManiFool Augmentation Crafting

A critical implementation detail is that for the crafting of ManiFool Augmentation images, black-box state-of-the-art models were used as the differential classifier f described above. Those models were previously trained on the given datasets but are not included in the evaluation phase of this work to avoid bias and to guarantee that the datasets are previously unseen by all the models that are being evaluated.

Tab. 4.1. Comparative evaluation of models trained on Dermofit using different augmentation methods and ManiFool Augmentation. Table published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

		None	Random	Erasing	ManiFool
ResNet	Original Test	0.7379	0.7859	0.7867	**0.8126**
	Random Affine	0.6515	0.6962	0.6573	**0.7900**
	Random Projective	0.4373	0.4817	0.4555	**0.6263**
VGG	Original Test	0.7526	0.8080	0.7924	**0.8258**
	Random Affine	0.6993	0.7387	0.6751	**0.8011**
	Random Projective	0.4319	0.5140	0.5071	**0.6200**
Inception	Original Test	0.7303	0.8051	0.7898	**0.8275**
	Random Affine	0.5544	0.7063	0.7123	**0.7883**
	Random Projective	0.2149	0.4388	0.4630	**0.5376**

4.4.2 Results and Discussion

The results of the ablative experiments, as well as baseline comparisons, will be addressed in this Section, as will the effects of the proposed method on the models' performance and robustness.

Performance improvement with ManiFool Augmentation

Tables 4.1 and 4.2 showcase the results of the ablative and baseline evaluation of the proposed ManiFool Augmentation method for the Dermofit and DDSM. First, we can see that the performance of models trained without any augmentation is substantially lower due to overfitting and limited manifold exploration. Random Augmentation improves the model performance. However, it offers no guarantee about the increase in the variance the model is exposed to during the training phase. Moreover, random augmentation can create out-of-distribution samples, which could obstruct model training.

Augmented images crafted by ManiFool originate from the same distribution as the original training data, a trait particularly critical in the setting of medical applications, where misclassifications can have undesired clinical effects. Moreover, Manifool Augmentation, due to its increased exploration capabilities, improves the accuracy by $2\% - 3\%$ for both datasets and all model architectures. Furthermore, ManiFool Augmentation consistently outperforms Random Erasing, Random Augmentation, and GAN Augmentation by approximately 2% for all datasets and models.

Limitations of Augmentation with GANs

Generating synthetic images with GANs is a task widely investigated, as was discussed in the Related Work Section. However, there are limitations regarding GANs for medical imaging applications: For many cases, the synthetic images suffer from low resolution, leading to a

Tab. 4.2. Comparative evaluation of models trained on DDSM using various augmentation methods and ManiFool Augmentation. Table published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

		None	Random	Erasing	DCGAN	ManiFool
ResNet	Original Test	0.8321	0.8254	0.8294	0.8228	**0.8426**
	Random Affine	0.7225	0.6849	0.6073	0.6964	**0.7970**
	Random Projective	0.2483	0.2078	0.3245	0.2657	**0.3245**
VGG	Original Test	0.7914	0.8381	0.8377	0.8405	**0.8443**
	Random Affine	0.2444	0.6547	0.7194	0.7371	**0.8094**
	Random Projective	0.1901	0.2046	0.2388	0.2279	**0.2733**
Inception	Original Test	0.8438	**0.8454**	0.8424	0.8414	0.8451
	Random Affine	0.4854	0.6423	0.6006	0.6980	**0.7330**
	Random Projective	0.1954	0.2164	0.2019	0.1980	**0.2356**

significant loss of information and quality. Additionally, GANs trained on the entire dataset do not provide ground truth labels for the generated samples. Therefore in order to use synthetic images as data augmentation with their respective label, we have to train n conditional GANs [225], where n represents the number of classes. This is not only time-consuming but also sometimes unachievable due to limited data.

For instance, some classes of the Dermofit dataset only have 23 images for training, which are not enough training samples for a conditional GAN. Efforts have been made to overcome the GAN labeling problem for medical imaging [33], by generating Brain CT scans along with paired segmentation label maps. However, this method does not guarantee the correctness of the label maps, and while the performance improvement on the test set is promising, mislabeling could cause ambiguity during training and jeopardize the model's robustness.

Furthermore, compared to Manifool Augmentation, augmentation with GANs does not necessarily increase the variance of the training data since images are sampled randomly from the training distribution and not from the outer regions of the manifold as achieved by ManiFool and can be seen in Fig. 4.1.

Robustness to Random Geometric Transformations

A noteworthy finding highlighted in Tables 4.1 and 4.2 is the substantial improvement in the robustness of models trained with ManiFool Augmentation to random transformations. The improvement is not only significant because it ranges from 7% to 15%, but also because, despite the fact that the proposed augmentation only used affine transformations, the robustness to projective transformations was greatly improved as well. The remaining evaluated augmentation methods, namely Random Erasing and GAN augmentation, provided marginal to no improvement in the robustness of the networks compared with standard random augmentation.

Fig. 4.3 depicts another experiment evaluating the effect of ManiFool Augmentation on the robustness of trained models. As described in the Methodology Section, Equation 4.9

Tab. 4.3. Comparative evaluation of models trained on Dermofit with various augmentation techniques and deployed on HAM10k, a previously unseen skin lesion classification dataset. Table published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

	None		Random		Erasing		ManiFool	
	Dermofit	**HAM10k**	**Dermofit**	**HAM10k**	**Dermofit**	**HAM10k**	**Dermofit**	**HAM10k**
ResNet	0.7379	0.1983	0.7859	0.3847	0.7867	0.1699	**0.8136**	**0.3854**
VGG	0.7526	0.1911	0.8080	0.3101	0.7924	0.1947	**0.8238**	**0.3419**
Inception	0.7303	0.2798	0.8051	0.2520	0.7898	0.2140	**0.8275**	**0.3009**

Tab. 4.4. Reported average robustness measure score defined in Equation 4.8 for different commonly used model architectures. Table published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

	ResNet	VGG	Inception
Dermofit	2.128	2.660	**3.391**
DDSM	**1.510**	1.240	1.242

measures the misclassification rate of a classifier for images transformed with random affine transformations for a specific range of geodesic distances.

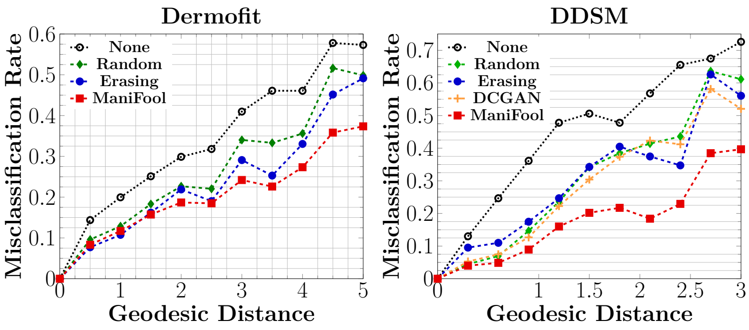


Fig. 4.3. Robustness of models with various augmentation approaches to random transformations with increasing geodesic distance. Figure published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

In Fig. 4.4 we show examples generated within a range of $G \in [1, 5]$ for Dermofit and $G \in [1, 3]$ for DDSM that were used to infer the misclassification rates of the trained models. As can be observed in Fig. 4.3, the models trained with ManiFool Augmentation had substantially lower misclassification rates for higher values of geodesic distance G.

Effect on Cross-Dataset Performance

To demonstrate the improved robustness achieved by ManiFool Augmentation, we perform cross-dataset evaluation between Dermofit and HAM10000 [284], which consists of 10.000 skin lesion images with 7 overlapping classes among the two datasets. Notably, all models trained with ManiFool Augmentation achieve $1\% - 5\%$ higher accuracy on the unseen dataset, as shown in Table 4.3. This validates the hypothesis that ManiFool Augmentation increases

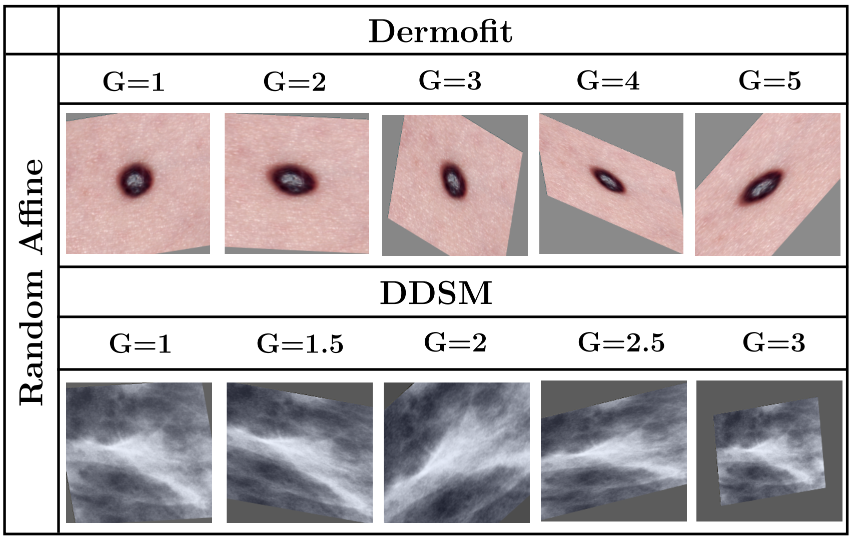


Fig. 4.4. Images generated with Random Affine Transformations for Dermofit [18] and DDSM [119] for a given range of Geodesic Distances G. Figure published in [217], used with permission from Springer Nature Customer Service Centre GmbH.

the model's understanding of the underlying data distribution, resulting in increased model robustness not only on geometric transformations but also on unseen test samples.

Robustness of Different Architectures

After we employ a classifier f to create ManiFool Augmentation samples, we can compute the average geodesic distance between the original and transformed images (Equation 4.8). This measure is able to quantify the robustness of a machine learning model since it implicitly calculates the distance between the decision boundaries of the trained model. Thus, models with larger geodesic distances between classes will be characterized by higher robustness. Other works [215] attempted to evaluate the robustness of a classifier using adversarial examples. However, such images cannot appear naturally, and no quantitative measures had been given for a model's robustness. After we crafted samples with ManiFool Augmentation, we inferred the robustness scores for the trained classifiers, and the results are shown in Table 4.4. This experiment underlines how the robustness of various architectures can differ according to each dataset. Thus, using a state-of-the-art architecture based on its results on an independent dataset is insufficient because its robustness can vary significantly. In our experiments, we found that InceptionV3 was the most robust model for Dermofit, while ResNet18 was the most robust for DDSM.

Training Dynamics Improvement 5

In this Chapter we describe the second contribution of this dissertation which has been published in [216]. Figures 5.2-5.5 and Tables 5.1-5.3 are used with permission from Springer Nature Customer Service Centre GmbH with License Number: 5058290012947.

5.1 Introduction

Fully Convolutional Neural Networks (F-CNNs) have been integrated into many Computer Assisted Diagnosis (CAD) systems, performing a plethora of medical image analysis tasks with increased complexity and requirements [192]. As a result, their complexity has grown, reaching hundreds of layers and millions of trainable parameters.

Another factor contributing to the size explosion of F-CNNs in CADs is that medical data is typically volumetric in nature and has been steadily increasing in resolution. This increase in the parameters of F-CNNs has drawbacks in terms of computation, energy usage, and storage requirements. To this end, we describe, for the first time in 3D F-CNNs, a quantization-based method that achieves model compression without any loss in performance and can decrease overfitting when training on limited datasets.

Various state-of-the-art segmentation techniques process volumetric data slice-wise leveraging 2D F-CNNs [237]. Such approaches achieve satisfying results but are not capable of fully exploiting the contextual information from neighboring slices. 3D F-CNNs, such as V-Net [198], 3D U-Net [65] and VoxResNet [55] have reached state-of-the-art performance in various segmentation applications using 3D convolutions kernels. However, their millions of trainable parameters, even though they increase the model capacity, require a large amount of training data and storage space for the model weights.

It has been shown [276] that deep learning-powered CADs using 3D F-CNNs have started being deployed into the medical workflow. Hospital infrastructure systems were already burdened with storing large medical patient records but now have to allocate further storage for trained model weights used in CAD systems. Moreover, the evolving area of patient-specific treatment [287], will increase the need for storage in medical facilities even more and enhance personalized diagnosis and monitoring.

Neural network compression has been an active field of research focusing on integrating state-of-the-art F-CNNs in low-power and resource-limited devices, such as smartphones and embedded electronics. A further possible use-case is represented by the decentralized training scheme of Federated Learning [161]. Even if client data is kept private, iteratively sending millions of parameters over the internet could be hindered by unstable connections; thus, compressed models could be a suitable approach for global training.

Contributions:

- We introduce, for the first time in 3D F-CNNs, a quantization approach with a novel bit-scaling scheme which we call 3DQ. 3DQ employs two trainable scaling factors and a normalization parameter that enhances the learning capacity of the model while maintaining compression.
- We thoroughly validate 3DQ on the critical task of 3D whole-brain segmentation, highlighting that our proposed approach can achieve state-of-the-art performance and impressive compression rates.
- We show that network quantization can improve training dynamics of large networks trained with limited data and achieve less overfitting.

5.2 Related Work

5.2.1 Model Quantization

Various approaches, such as parameter pruning, low-rank factorization, knowledge distillation, and weight quantization [60] have been proposed to compress the size of CNNs without compromising their performance. Particularly, weight quantization to binary [231] and ternary values [121, 176, 312] has been widely investigated for various applications due to its additional benefit of allowing for impressive speed-up during both training and inference by approximating convolutions with XNOR and bitcounting operations [231]. Even though it has been shown that XNOR-Net revolutionized this speed-up [231], there is also a significant trade-off between accuracy and speed, which is not ideal for medical applications.

Chen et al. [57] proposed Layer-wise/Limited training data Deep Neural Network Quantization (L-DNQ), where they formulated quantization for each layer as a discrete optimization problem.

TernaryNet [121] was the first approach in medical imaging to propose compact and fast F-CNNs utilizing ternary weights, where a 2D U-Net was used for slice-wise pancreas CT segmentation.

5.2.2 Robustness of Quantized Models

The robustness of quantized neural networks has also been explored. Guo et al. [109] were the first to show the intrinsic relationship between weight sparsity and DNN robustness against FGSM and DeepFool adversarial examples. They conducted experiments by pruning both the weights and the activations of DNNs and concluded that appropriately higher model sparsity could lead to better robustness for nonlinear DNNs.

Xiao et al. [299] showed that weight sparsity is also beneficial for network robustness verification. They demonstrated that weight sparsity could turn computationally intractable verification problems into tractable ones and improved weight sparsity in DNNs by training them with L_1 regularization. Additionally, weight sparsity significantly speeds up the linear programming solvers [146] for network robustness verification [233].

Feng et al. [85] proposed an adversarial attack against the Deep product quantization network (DPQN) [47], that performs fast image retrieval tasks on large-scale datasets. The proposed attack is called product quantization adversarial generation (PQ-AG) and can lead DPQN to produce semantically irrelevant results. Guo et al. [108] proposed an attack and a defense mechanism for quantized networks that is based on adversarial training. They developed an Iterative Quantized Local Search (IQLS) algorithm that computes strong perturbations by quantizing both input and perturbation space. Afterwards, they introduce an efficient Quantized Adversarial Training (QAT) scheme based on the upper bound of iterations needed for IQLS. Khalid et al. [151] proposed Constant Quantization (CQ) and Trainable Quantization (TQ) to enhance the robustness of DNNs against adversarial examples. CQ quantizes input pixels based on a specified number of quantization levels, while TQ learns the quantization levels iteratively during training to further increase the defense's strength.

A boundary-based retraining method was proposed in [260] combining adversarial and quantization losses and adopting a nonlinear mapping scheme to defend against white-box gradient-based adversarial attacks. They experimentally showed that their method could retain its accuracy after quantization better than other baselines on black-box and white-box adversarial attacks. Yoon et al. [308] introduced Stochastic Quantized Activation (SQA) that tackles overfitting issues and achieves robustness combined with FGSM adversarial training even against white-box attacks. SQA reduces the adversarial effect by providing random selectivity to activation functions.

Lin et al. [176] introduced Defensive Quantization (DQ) which controlled the Lipschitz constant of the DNN during quantization so that the magnitude of the adversarial perturbation would remain non-expansive during inference time. The novel approach outperformed the full precision DNNs in terms of robustness and hardware efficiency. To combat the challenge of data scarcity, Choi et al. [63] employed a data-free quantization approach, which was combined with knowledge distillation [125]. Their approach is shown in Fig. 5.1. First, they

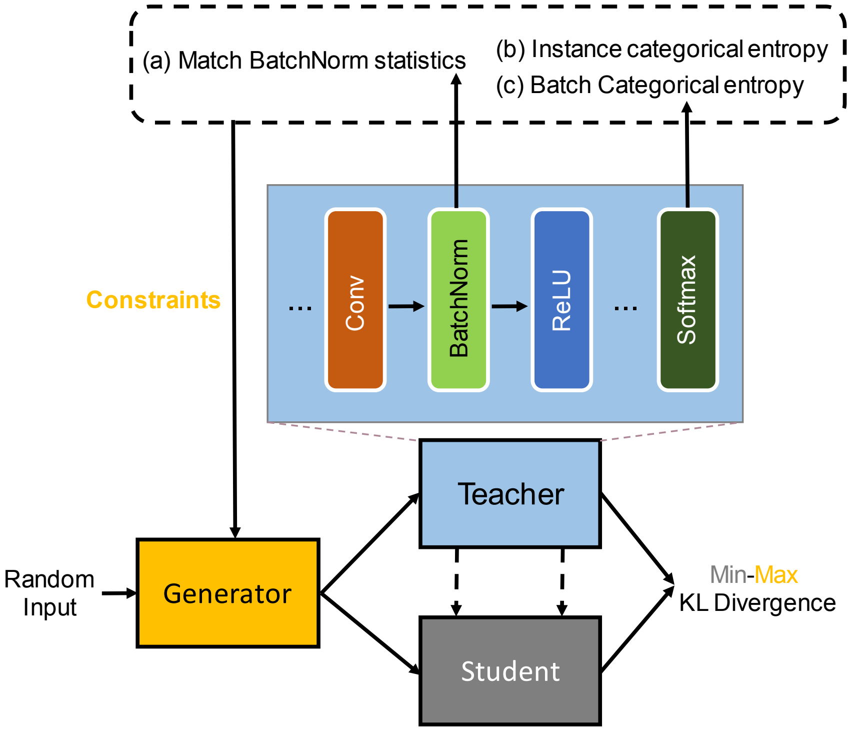


Fig. 5.1. Overview of Data Free Defensive Distillation proposed in [63]. For the adversarial example generation the statistics from the batch normalization layers were matched between the generated and the original data using KL divergence and no other data were used from the training set.

minimized the maximum distance between the outputs of the teacher and the quantized student for adversarial samples crafted by a generator. For the adversarial example generation, instead of using the original data, the statistics from the batch normalization layers were matched between the generated and the original data using KL divergence. Their data-free quantized models achieved comparable performance to models trained on the original large-scale datasets.

5.3 Methodology

5.3.1 Weight Quantization

The primary goal of our quantization scheme is approximating the full precision weights of a 3D convolutional model W by their ternary counterparts {-1, 0, 1}, $\tilde{W}$, as shown in Eq. 5.1.

The first step is inferring threshold Δ, based on which W will be categorized into three quantization bins. Other methods use a single Δ for the entire model [121]. However, 3DQ calculates one Δ per layer y, in order to maintain the variability in the range of weight values within each layer and overcome weight sparsity [312]. Specifically, each Δ_l is calculated as $\Delta_l = t \cdot \max(|W_l|)$: the maximum absolute value of weights in each layer is multiplied by a constant hyperparameter t which modulates weight sparsity and remains consistent for all

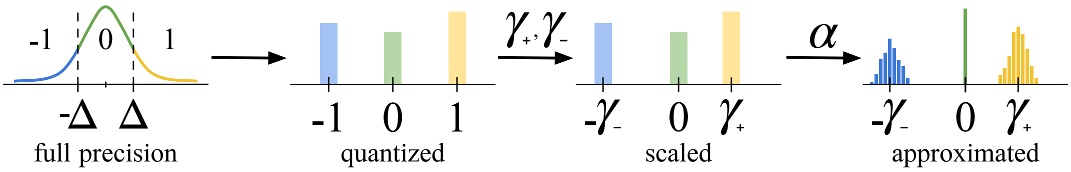


Fig. 5.2. Overview of 3DQ. The full precision weights are quantized into ternary values, scaled by $\gamma^{\pm}$ and then further diversified by the factor α. Figure published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

layers. t has been set to 0.05, following [312] since it achieved the optimal trade-off between weight sparsity and accuracy.

Training an entire F-CNN with ternary weight values {-1, 0, 1} would lead to suboptimal performance. Therefore, after thresholding, the computed ternary weights $\tilde{W}$ are multiplied by a set of scaling factors. 3DQ uses two scaling factors, γ_l^+ and γ_l^- [312], which are trainable parameters learned for each layer l, differing from previous methods [121, 231].

$$W_l \approx \tilde{W}_l = \begin{cases} +\gamma_l^+ \cdot \alpha & \text{if } \; W_l > \Delta_l \\ 0 & \text{if } \; |W_l| < \Delta_l \\ -\gamma_l^- \cdot \alpha & \text{otherwise.} \end{cases} \tag{5.1}$$

Moreover, unlike [312], we introduced one more scaling factor α to 3DQ [83]. α is calculated from W as the average of the model weights with an absolute value larger than Δ_l,

$$\alpha = \frac{1}{n_{\Delta_l}} \sum |\tilde{W}_l||W_l| \tag{5.2}$$

where $n_{\Delta_l} = \sum |\tilde{W}_l|$. α improves the approximation of the full precision weights, because it spreads the quantized weight values within the same bin, leading to increased expressivity and diversity in the weights between the channels of each layer. The proposed quantization pipeline is shown in Fig. 5.2 and our method has been published in [216].

5.3.2 Storing Compressed Weights

Full precision W and ternary $\tilde{W}$ are both required during training in order to perform model optimization and learn the scaling factors γ^+ and γ^-. However, during inference, the full precision weights W are no longer required, and thus there is no need to store them.

After scaling the ternary weights $\tilde{W}$, the values still take up 32 bits; therefore, it is crucial to store the model in a way that ensures the 16x compression rate that the ternary weights can achieve. To this end, we split each kernel into three components, as shown in Fig. 5.3: 1) A pair of learned scaling factors γ^+ and γ^- for each layer of a F-CNN architecture. 2) The values of α, which are inferred from the full precision weights and sum up to as many as the channels of each layer. 3) The ternary weights, which make up most of the model parameters, and can sum up to millions of values for large 3D models.

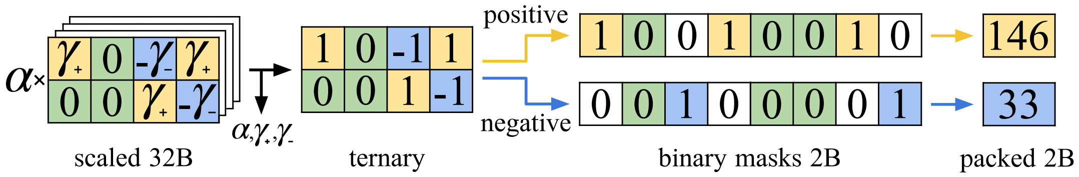


Fig. 5.3. Overview of 3DQ compression technique for storing the model parameters. The scaling factors are split from the ternary weights, which are further separated into two binary vectors, then packed from 8 bits to 1 byte. The same process is followed backwards to recover the stored weights of a trained model. Figure published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

The scaling factors are stored as full precision variables that take up 32 bits of disk space each. In the meantime, each ternary weight kernel is divided into two binary masks, one for positive weights and the other for negative weights. Unmarked areas in both masks represent zero weights. The masks undergo bit packing, and 8 weight bits are stored in 1 byte. The same procedure is followed backwards to load the compressed saved models: first, unpack the ternary weight values, then multiply them with the stored full precision scaling factors. The described technique achieves high compression rates, which is critical for large 3D networks, which require storing up to 45M parameters for each model [198].

5.4 Experiments

5.4.1 Experimental Setup

Datasets

We validated 3DQ on two publicly available medical imaging 3D segmentation datasets, the Multi-Atlas Labelling Challenge (MALC) [186] and the Hippocampus (HC) Segmentation dataset from the Medical Decathlon challenge [257]. MALC belongs to the OASIS dataset and consists of 30 whole-brain MRI T1 scans with manual expert annotations. The input volumes are sized $256 \times 256 \times 256$, which were sampled in cubic patches of size $64 \times 64 \times 64$. Maintaining the original challenge split, we leveraged 15 scans for training and 15 for testing. We considered 28 classes for the segmentation, as in [237], and we repeated all the experiments 5 times with different initialization seeds.

HC includes 263 training samples with average size $36 \times 50 \times 35$, which we padded to cubes sized $64 \times 64 \times 64$. The challenge test set is not available to the public; thus, we performed 5-fold cross-validation, dividing the dataset to 80% for training and 20% for testing using patient-level splits. For HC, the voxels are categorized into 3 classes, including 2 parts of the hippocampus (hippocampus proper and hippocampal formation) and the background [257].

Model Training

To showcase the generalizability of our method, we quantized commonly used 3D F-CNN architectures, specifically 3D U-Net [65] on MALC and HC, and V-Net [198] on MALC. The

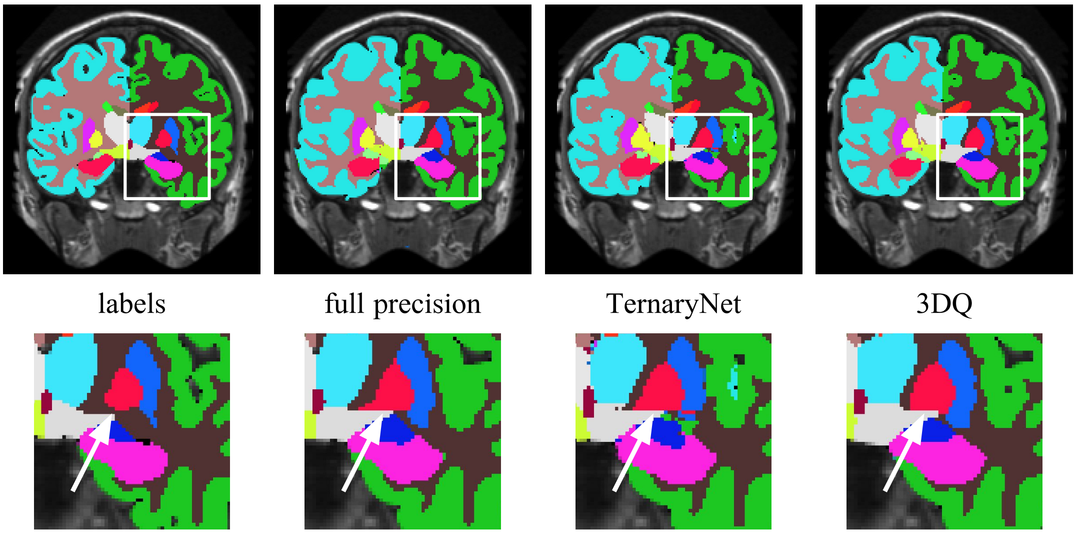


Fig. 5.4. Qualitative results of 3DQ in comparison with various baseline methods. White arrows on the zoomed views highlight the superior segmentation performance achieved by 3DQ. Figure published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

utilized models are suitable candidates for quantization and compression since they consist of 16M and 45M trainable parameters, respectively, and take up to 175MB to store on the disc.

For both MALC and HC, we trained the models with an equally balanced loss function combining Dice loss and weighted cross-entropy to alleviate class imbalance. The class weights were calculated using median frequency balancing [238]. Adam optimizer [153] was used, while the initial learning rate for 3D-UNet was 0.0001 and for V-Net 0.00005. All models were trained on an NVIDIA Titan Xp Pascal GPU and implemented on the deep learning framework PyTorch [218]. Even though all 3D U-Net models were trained from scratch, initiating the quantized experiments on V-Net with pre-trained weights was beneficial.

Evaluation Metrics

Since we aim to compress quantized models without losing performance compared to the full precision models, we evaluate our method based on two different criteria: the Dice Score achieved by the networks across MRI volumes and the storage space required to save the models on the disc. We report the average Dice Score across the 5-folds in case of HC or 5 repetitions for MALC and the corresponding standard deviation.

Ablative testing

To evaluate the effectiveness of the primary components of 3DQ, specifically the ternary weights and the incorporation of scaling factor α, we performed ablative testing. We compared 3DQ with BTQ (Binary Trained Quantization), an adjusted binarized version of Trained Ternary Quantization (TTQ) [312], to underline the advantages of ternary weights. Moreover, we compared 3DQ against TTQ to showcase the contribution of scaling factor α to the model performance.

Baseline comparison

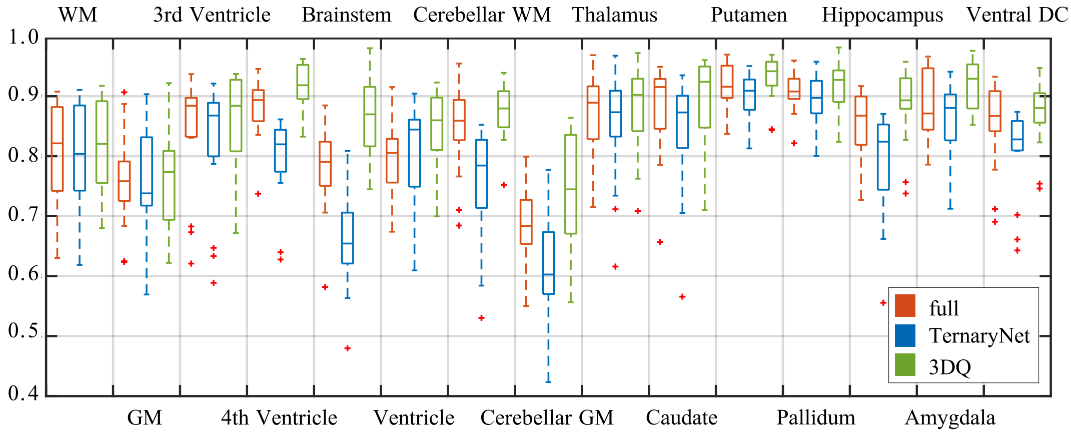


Fig. 5.5. Box-plot of Dice scores of 3D U-Net comparing 3DQ with its full precision counterpart and TernaryNet on the classes of the right hemisphere of the brain for the 15 testing volumes of MALC [186]. Figure published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

Tab. 5.1. Comparison of Dice scores of 3DQ with TTQ and its binarized version BTQ on HC and MALC, with 3D U-Net and V-Net. ± denotes the standard deviation. Table published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

		BTQ	TTQ	3DQ
HC	3DU-Net	0.847 ± 0.009	0.912 ± 0.008	**0.915** ± **0.006**
MALC	VNet	0.770 ± 0.013	0.790 ± 0.010	**0.802** ± **0.004**
	3DU-Net	0.735 ± 0.005	0.828 ± 0.007	**0.844** ± **0.006**

3DQ was compared with its full precision counterpart to evaluate whether quantized networks are capable of matching the performance of full precision models. TernaryNet [121], which was recently proposed for the compression of 2D U-Net [236] for pancreas CT segmentation was also compared with 3DQ. As an alternative compression baseline, we selected knowledge distillation with a temperature $T = 40$ to train scaled-down versions [125] of 3D U-Net and V-Net, that take up exactly the same storage space as the quantized networks compressed with 3DQ.

5.4.2 Results and Discussion

Ablative Testing

Table 5.1 shows that models quantized with ternary weights outperform their binary versions by 3-11% for MALC and 7% for HC due to their higher learning capacity justifying the choice of ternary weights in 3DQ. Furthermore, Table 5.1 highlights the positive effect of scaling factor α, which allowed the performance of 3DQ to be increased by 1-2% in comparison to TTQ for both datasets, with lower standard deviation. α overcomes the quantization drawbacks since it enables the ternary weights to have a larger range of values, better approximating their full precision counterparts.

Comparative Methods

As can be observed in Table 5.2, 3D U-Net quantized with 3DQ performs 1% better than the full precision model in the case of HC and over 2% better for MALC. This can be attributed to the fact that model quantization acts as a regularization method by limiting the dynamic range of the weights. Specifically, in the task at hand, where the models were trained with a limited amount of volumes, reducing the model capacity with the quantized weights leads to reduced overfitting and better generalization for 3D U-Net.

3DQ also outperformed TernaryNet in all experiments for MALC and HC, with a margin ranging from 7 to over 10%. We attribute this to the learned scaling factors $\gamma^{\pm}$ and the absence of the hyperbolic ternary tangent that clips the activation values and limits the network's capacity.

Figure 5.4 showcases sample segmentations for a slice of volume from MALC with prediction maps produced for 3D U-Net. A zoomed-in view of the segmentation maps highlights crucial subcortical structures with a white arrow. The full precision and TernaryNet predictions are characterized by over-inclusions of small structures and misclassified areas. The box plot in Figure 5.5 certifies the higher quality of the segmentation maps predicted by models quantized with 3DQ, showing the Dice scores on the right hemisphere structures. 3DQ performed better than both full precision and TernaryNet, and had fewer outliers, exhibiting more uniform results among all test samples.

Comparison with Knowledge Distillation

Another experiment shown in Table 5.2 is comparing 3DQ with knowledge distillation. The distilled networks have 16x fewer parameters than the full precision models in order to match the 3DQ model sizes while retaining full precision weights. Despite the fact that the compressed networks achieve almost equal performance with the full prediction model for HC, the margin is increased for MALC, where the student distilled models achieved 9-10% lower Dice score than the full models for both 3D U-Net and V-Net. This drop in Dice score can be attributed to the 16x smaller size of the distilled models in comparison to the original ones. Additionally, the student networks rely on a teacher model's predictions, limiting their learning capacity [125]. 3DQ is a successful model compression technique since it outperforms the distilled networks in all cases by a substantial 8-11% on MALC.

Quantization on Different Architectures

Table 5.2 shows the impact of quantization in two different 3D model architectures, namely 3D U-Net and V-Net. Even though 3D U-Net is 3x smaller than V-Net, it reached higher Dice scores in our experiments on MALC, especially after quantization. While the full precision networks achieved similar Dice scores with a difference of 1%, the quantized 3D U-Net achieved 4% higher Dice score than the quantized V-Net. This difference in performance can be attributed to the fact that MALC consists of only 15 training volumes, which is a very small amount of data to train V-Net, which has 45M parameters compared to 3D U-Net that has 16M. Therefore 3D U-Net was a more suitable model for the task at hand and achieved less overfitting and improved generalizability for both datasets and all baselines.

Tab. 5.2. Comparison of Dice scores of 3DQ with baseline methods. Tests performed on HC and MALC, with 3D U-Net and V-Net. ± denotes the standard deviation. Table published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

		Full	Distilled	TernaryNet	3DQ
HC	3DU-Net	0.914 ± 0.005	0.908 ± 0.019	0.845 ± 0.013	**0.915 ± 0.006**
MALC	VNet	**0.815 ± 0.008**	0.715 ± 0.001	0.696 ± 0.016	0.802 ± 0.004
	3DU-Net	0.822 ± 0.005	0.730 ± 0.008	0.774 ± 0.012	**0.844 ± 0.006**

Tab. 5.3. Model size in MBytes for full precision and baseline compressed models. Table published in [216], used with permission from Springer Nature Customer Service Centre GmbH.

	Full	Distilled	Ternary	Binary
3DU-Net	63MB	3.9MB	3.9MB	**2.0MB**
V-Net	175MB	11MB	11MB	**5.5MB**

Compression

The storage size for the models is showcased in Table 5.3. Quantized ternary weights in TernaryNet, TTQ, and 3DQ reduce the storage requirements by a factor of 16, compared to full precision models. The introduced scaling factors influence the storage by only a few KBytes. Binary weights further reduce the storage size by 2 times compared to the ternary ones, at the cost of decreased segmentation performance, due to limited training capacity.

Part III

Robustness Evaluation

6 Model Benchmarking

In this Chapter we describe the third contribution of this dissertation which has been published in [215]. Figures 6.5-6.7 and Tables 6.1-6.3 are used with permission from Springer Nature Customer Service Centre GmbH with License Number: 5058281246956.

6.1 Introduction

Thorough model evaluation is crucial before integrating a DNN into a framework deployed in the real world. Applications such as autonomous driving, online financial systems, and, of course, healthcare deal with sensitive input data and are required to make critical decisions. Testing a model's performance on an unseen test set is the most commonly used practice, which can successfully measure its performance on data points from the same distribution as the training set. However, a machine learning-powered system could be subjected to various unexpected input types, from outliers, unseen classes, edge-case samples, or, as we discussed previously, adversarial examples.

To that end, developing thorough model benchmarking techniques is valuable and necessary for model deployment.

6.2 Related Work

6.2.1 Benchmarking Datasets

A straightforward way to evaluate a model's performance under many circumstances is standardized, publicly available benchmarking datasets investigating various types of adversity.

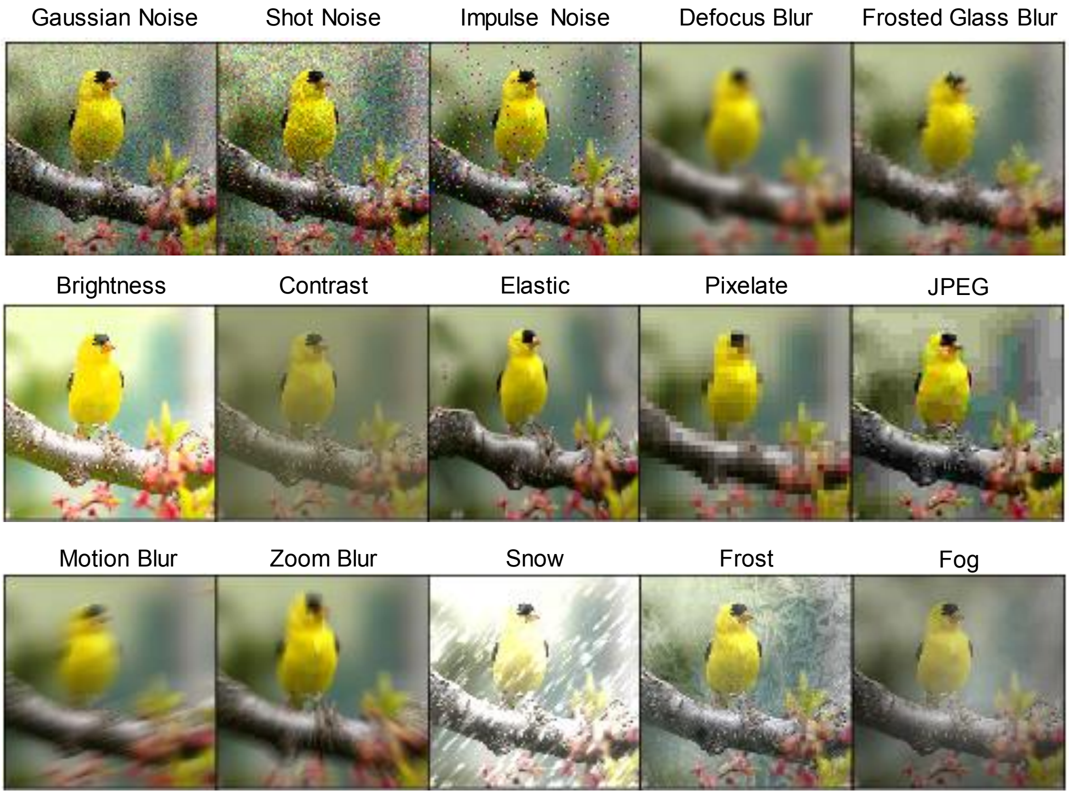


Fig. 6.1. Examples produced with the distortion mechanisms proposed by Hendryks et al. [123]. The images are part of the ImageNet-C Benchmarking dataset [74, 123].

Corruptions

Hendrycks and Dietterich [123] introduced two benchmarks, ImageNet-C, which includes standardized corruption types, and ImageNet-P, which benchmarks a classifier's robustness to common perturbations. In Fig. 6.1 some examples are shown from ImageNet-C along with their corresponding corruption type. ImageNet-C, includes 15 common visual corruptions applied to the ImageNet dataset [74].

Gaussian noise can appear in low-lighting conditions, *shot or Poisson noise*, is electronic noise caused by properties of light. *Impulse noise* is a color analog of salt-and-pepper noise caused by bit errors. *Defocus blur* happens when a sample is out of focus. *Frosted Glass Blur* simulates frosted glass windows. *Motion blur* occurs when a camera is moving fast. *Zoom blur* happens when a camera moves toward an object quickly. *Snow* is a visually obstruction caused by weather phenomenons. *Frost* simulates the lenses or windows when they are coated with ice crystals. *Brightness* simulates varying daylight intensity. *Contrast* depends on the lighting, and the color of the pictures objected. *Elastic transformations* stretch or contract small regions in the image. *Pixelation* is an effect of upsampling or downsampling an image. Finally, *JPEG* is an image compression format that could introduce compression artifacts [123].

The corrupted samples should only be used during the inference of a model. To evaluate a model's robustness to a specific type of corruption, they also introduced a score across five corruption severity levels. The aggregated score across severity levels is called Corruption

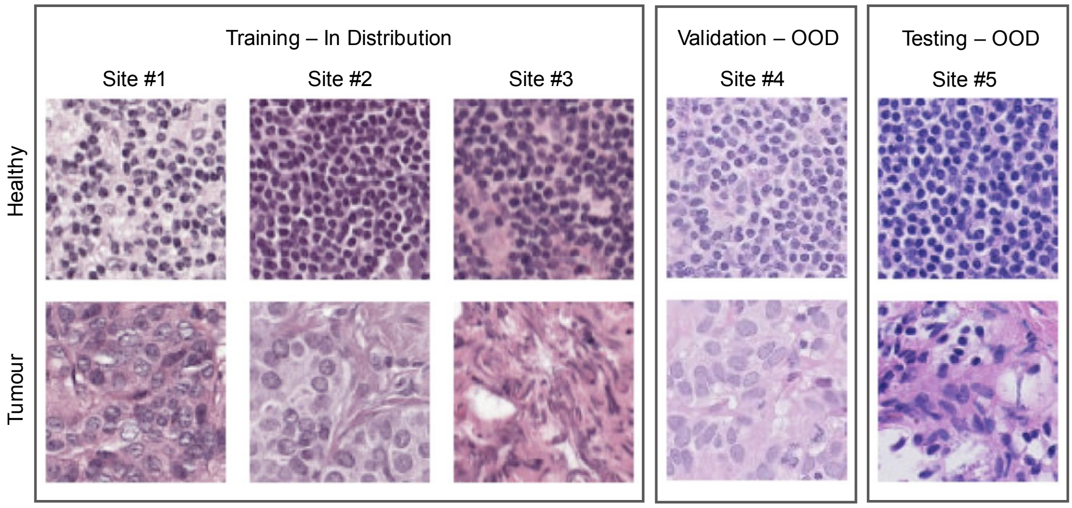


Fig. 6.2. Examples taken from the Camelyon17-Wilds Benchmarking dataset in [159] The aim is to predict the presence of tumor tissue in patches taken from sites that are not included in the training set. Patches released in the Chamelyon 17 dataset [22].

Error [123]. Their experiments found that even state-of-the-art architectures are susceptible to data corruption and that defense strategies should be employed to increase their robustness.

Hendrycks et al. [122] also introduced ImageNet-Renditions (ImageNet-R), a test set of 30,000 images with renditions i.e. paintings, embroidery, and more of ImageNet [74]. These variations are naturally occurring, with textures and local image statistics. This benchmark can be used to investigate whether a classifier suffers from texture bias [95] or is biased to synthetic images [274]. They also explored natural shifts within the image capturing process and introduce a benchmark that includes object occlusion, a shift in orientation, zoom, and scale at test time [122].

Tian et al. [279] introduced DeepTest, a systematic testing tool for detecting erroneous behaviors for autonomous driving powered by DNNs. The method synthesizes test cases that maximize neuron coverage [219]. Neuron coverage is the ratio of unique neurons that get activated for an input over the total number of neurons in a DNN. The synthetic test set includes the following distortions: varying brightness, Contrast, translation, scaling, horizontal shearing, rotation, blurring, and adding fog and rain [279]. Those transformations are applied to an input to maximize the neuron coverage. Experiments showed that neuron coverage varied significantly for different input-output samples and different types of transformations. Thus, a neuron-coverage-based testing scheme could help in identifying the edge cases.

Kaman and Rother [142] benchmarked DNNs for the task of semantic segmentation against real-world corruptions. Their work extended the corruptions introduced in ImageNet-C. Experiments with a variety of datasets and architectures showed that Xception-based networks [64] were generally more robust than ResNets [117] and that MobileNet-V2 [243] was vulnerable to most image corruptions besides blurring [142].

Distribution Shift

Distribution shift should also be thoroughly evaluated for systems deployed in production. However, such shifts are usually not modeled in publicly available datasets. To this end, Koh et al. [159] recently introduced a collection of 8 benchmark datasets that include an extensive range of distribution shifts that naturally occur in real-world applications, such as variations across hospital sites, cameras for wildlife monitoring, and time or location in satellite imaging. For each dataset, they showed that standard training results in substantially lower out-of-distribution performance and that even if models are trained using methods to overcome the distribution shift, the gap remains. Specifically for medical imaging applications, they curated the Camelyon17 dataset [22]. Whole slide images from 3 hospital sites are used for training while images from different sites are used for validation and testing as out-of-distribution (OOD), as can be seen in Fig. 6.2.

The Breed dataset [245] is a large-scale subpopulation shift benchmark, in which the data subpopulations differ between training and evaluation. This dataset aims to assess how robustly models generalize beyond their training datasets. This approach leverages existing dataset labels to identify groups of semantically similar classes, called superclasses. For example, all different breeds of dogs belong to the superclass "Dog." Afterwards, the original dataset classes are considered subpopulations. The subpopulation shift is introduced by making the subpopulations in the training and test set disjoint. Thus, a model could be trained on the dog class "Dalmatian," but it would be evaluated on the class "Poodle." Their experiments found that model performance drops significantly on the shifted distribution and that models that are more accurate on the original distribution are often more robust to subpopulation shifts [245].

Adversarial Examples

Chet et al. [58] created DAmageNet, a dataset containing adversarial examples with a minimal perturbation but a high transfer rate among architectures. DAmageNet can be used as a benchmark to evaluate the robustness of DNNs to adversarial samples. DAmageNet consists of 50000 adversarial samples from ImageNet validation set crafted using Attack-on-Attention (AoA) also proposed in [58]. AoA attempts to change the attention heat map by shifting the attention away from the correct class. Experiments on 13 different commonly used model architectures showed that all models were vulnerable to DAmageNet, even when adversarial defenses were applied.

Similarly, Kang et al. [144] proposed an evaluation framework with Unforeseen Attacks called ImageNet-UA. They propose the following attack strategies: *JPEG attack*, where perturbations are computed in the compressed JPEG space, the *Fog attack* which is adversarially optimized, the *Snow attack* with adversarial perturbations that optimize its intensity and direction and the *Gabor noise attack*. Their experiments showed that even models that had been adversarially trained with attacks like PGD remain vulnerable to such distortion-based attacks.

Finally, RobustBench [69] contains an ensemble of white- and black-box attacks with clearly defined threat models that can be leveraged to evaluate the robustness of DNNs. The challenge examined various defense mechanisms, and pre-trained robust models are publicly available to be used for downstream tasks.

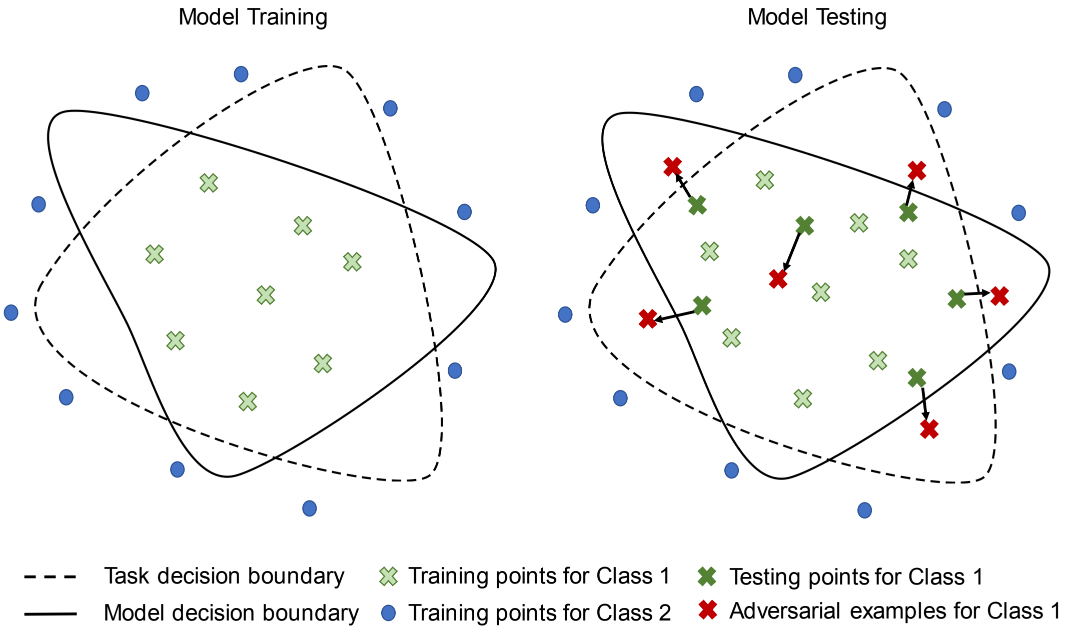

Model Training
Model Testing
Task decision boundary
Model decision boundary
Training points for Class 1
Training points for Class 2
Testing points for Class 1
Adversarial examples for Class 1

Fig. 6.3. Model Testing: The task and predicted model boundaries are shown, along with training, testing and adversarial samples. With model testing using adversarial examples we are creating inputs that are crossing the model's decision boundary and cause misclassification, uncovering the model's vulnerability to this type of input [277].

Medical Imaging

Recently, efforts have been made to benchmark different models regarding their robustness to outliers for medical applications. Specifically, the Medical Out Of Distribution (MOOD) Challenge [315] provided two large-scale standardized benchmark datasets, one consisting of brain MRI scans and one with abdominal CT scans. The training sets contain samples where no anomalies were identified. However, the test sets contain naturally occurring and synthetic anomalies to cover a broad and unpredictable range of outliers. The challenge aims to provide an analysis of the weaknesses and strengths of the methods based on various factors [315].

6.2.2 Verification Methods

The methods we discussed above perform *model testing*, i.e., evaluating a model under various conditions and monitoring its behavior. Testing can be performed on legitimate, "naturally occurring" inputs, or as we saw, it could include adversarial examples and other degenerate inputs. Even though testing can be sufficient for traditional machine learning applications, it does not provide security guarantees [277]. An attacker can still craft an input that was not seen in the test process, for example, a new adversarial attack or distortion.

An approach that can provide security guarantees for machine learning models is *model verification*. Creating guarantees that can define the space of inputs that are always processed correctly by a model is a critical step for adversarial security and model evaluation.

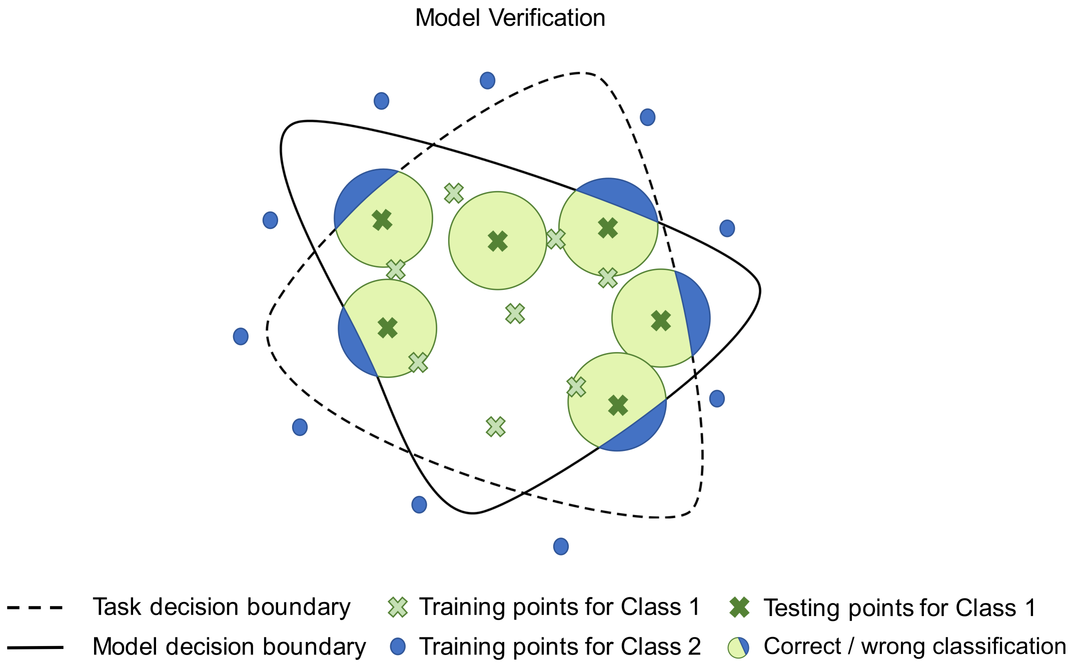

Model Verification
Task decision boundary
Model decision boundary
Training points for Class 1
Training points for Class 2
Testing points for Class 1
Correct / wrong classification

Fig. 6.4. Model verification. The task and predicted model boundaries are shown, along with training and testing samples. Model verification guarantees that a model's decision will be correct for a new sample for a constant region regardless of the type of the input [277].

Current approaches of model verification verify that a classifier f assigns the same class to every sample within a specified region around a sample x as can be seen in Fig. 6.4 [277]. Pulina et al. [224] proposed one of the first verification systems. They showed that the output class of a neural network is constant across a specified neighborhood. However, their system was limited to one hidden layer and to networks with few hidden units. Huang et al. [132] extended this method to deeper architectures that could be used for ImageNet [74] classification. Reluplex [146] is another verification system that uses linear programming solvers to scale to deeper architectures and specializes in ReLU networks. CLEVER [291] is an approach based on extreme value theory that can provide a lower bound on the minimal perturbation needed to generate an adversarial example. The proposed score is attack-agnostic and computationally feasible for deep architectures.

However, there are still limitations to the verification approaches, namely that the system relies on assumptions, such as that a given input is only relevant to a subset of the hidden neurons [277]. Thus, inputs that violate those assumptions can still harm the performance of verified models. Furthermore, they verify only that the output class remains constant in a given neighborhood of a sample x. However, it is challenging, if possible, to exhaustively validate all inputs x near which the prediction should remain constant. In the scope of this thesis, our contribution presented below is a model testing method.

6.3 Evaluating models for medical imaging applications with adversarial examples

6.3.1 Introduction

As previously discussed in this dissertation, deep learning is increasingly adopted within the medical imaging field for various tasks such as classification, segmentation, detection, and more. The traditional technique for the assessment of machine learning models consists of the evaluation of their *generalizability i.e.* their performance on unseen test cases. However, for applications with *limited* training data, such as medical imaging, training over-parameterized deep learning models can lead to the "memorization" of the training dataset. Validating the performance of such models on an available non-overlapping test set is common, yet substantially limited in exploring the model's robustness to outliers, noisy data or labels, and more. Furthermore, the limited interpretability of DNNs due to their "black-box" nature hinders their adoption into clinically-used frameworks.

Existing model benchmarking schemes focus on model over-fitting but insufficiently investigate cases of model sensitivity to variations of the input data. When a DNN is driven to its limits, robustness evaluation could estimate the likelihood of failure. In this work, we address model evaluation utilizing adversarial examples [273] that are purposefully crafted to fool a DNN and can uncover scenarios where its performance may decrease. Our method leveraging adversarial examples as a benchmark is also substantially less strenuous and costly than creating a sufficiently diverse test set with manual expert annotations.

Furthermore, creating synthetic distortions like brightness variations or weather condition changes is not directly applicable to medical applications. Modalities such as MRI or CT scans need to be carefully distorted, for example, with realistic acquisition or motion artefacts to create meaningful test samples. Shaw et al. [251], for instance, proposed an MRI data augmentation method introducing motion artefacts in k-space. Such an approach could increase the generalization and robustness of the model but could not be directly generalized to other medical imaging applications.

Furthermore, creating synthetic samples with GANs can lead to model hallucinations [67] since the training and target distributions do not always match and can create model bias. To that end, utilizing adversarial examples is a straightforward way to model edge cases and evaluate the robustness of DNNs for medical imaging applications.

As discussed earlier in this dissertation, adversarial examples are images created to fool machine learning models, while the added perturbations are not perceptible to human eyes [273], as can be seen in Fig. 6.5. Our work is among the first that investigate adversarial examples in medical imaging and use them in a constructive way to benchmark model performance not only on clean and noisy but also on adversarial data. In a medical setting, Zhu et al. augmented their dataset with adversarial examples to limit overfitting and increase the performance of their model on mass segmentation [314].

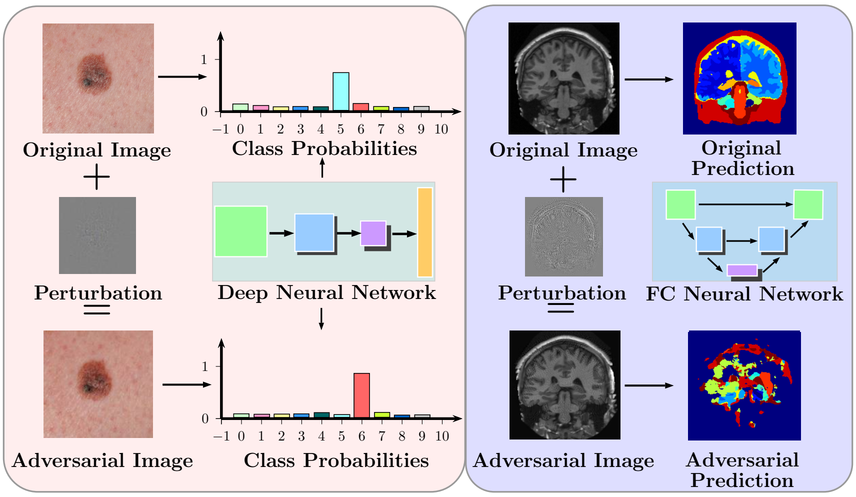


Fig. 6.5. Overview of Adversarial Crafting and its effect on the prediction of a DNN. Even though the difference between the crafted adversarial example and the original image is imperceptible, DNNs are successfully fooled into erroneous predictions. Figure published in [215], used with permission from Springer Nature Customer Service Centre GmbH.

Although adversarial examples may not occur naturally in acquired data, they could be used as benchmark during test time or during training to increase model robustness and optimize the decision boundaries learned for various tasks.

Contributions:

- We leverage adversarial examples crafted with 3 widely used attack mechanisms to benchmark the robustness of deep models.
- We highlight the difference between random noise and adversarial perturbation and show their different effect on the model embeddings and performance.
- We compare a variety of commonly used architectures, such as Inception [272] and UNet [236] for classification and segmentation and discuss the architectural features that contribute to their robustness or vulnerability.
- We demonstrate that widely used state-of-the-art models are not only vulnerable to adversarial examples but also exhibit substantially different behaviors under attack.

6.3.2 Methodology

Adversarial Crafting

For a trained model f and an original input image $\boldsymbol{x}$ with output label y we craft an adversarial image $\hat{\boldsymbol{x}}$ by solving a box-constrained optimization problem $\text{argmin}_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|$ subject to $f(\boldsymbol{x}) =$

y, $f(\hat{\boldsymbol{x}}) = \hat{y}$, $\hat{y} \neq y$. Such an optimization scheme minimizes the computed perturbation, $\boldsymbol{\epsilon}$, $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon}$, while simultaneously fooling the model f [273]. By enforcing a constraint such as $\|\boldsymbol{\epsilon}\| \leq \eta$, where η is a hyperparameter chosen by the attacker we can limit the added perturbation to be so small that it is imperceptible to human eyes.

Classification

Gradient-based adversarial example crafting techniques have been proposed to compute the minimum amount of perturbation $\boldsymbol{epsilon}$ that misclassifies $\hat{\boldsymbol{x}}$. Such techniques include the Fast Gradient Sign Method (FGSM) [102], DeepFool (DF) [202], Jacobian Saliency Map Attacks (SM) [213] and Projected Gradient Descent (PGD) [183]. Adversarial examples generated with some of these methods for dermatology and radiology are shown in Fig. 6.6. For a trained classifier f, FGSM [102] performs a one-step update along the sign of the gradient that maximizes the task loss $\mathcal{L}$ and the resulting perturbation is described as $\boldsymbol{\epsilon} = \eta \text{sign}\left(\nabla_{\boldsymbol{x}} \mathcal{L}(\theta, \boldsymbol{x}, y)\right)$, where θ are the parameters of the model. The strength of the added perturbation is determined by a hyper-parameter η that is in most cases assigned a low value so that $\hat{\boldsymbol{x}}$ is imperceptible from $\boldsymbol{x}$.

In contrast to FGSM, DeepFool [202] consists of an iterative greedy search process. In every iteration, the projections of the input image to the decision boundaries of all classes are calculated, and an $\boldsymbol{\epsilon}$ is inferred to push $\boldsymbol{x}$ towards the decision boundary of the closest class, besides the ground truth. In Saliency Map Attacks [213], we estimate the impact of each pixel on the prediction of the DNN, and afterwards, the input image is selectively perturbed so that it causes the highest-impact change to the prediction.

Segmentation

In [301], Dense Adversarial Generation (DAG) was proposed for crafting adversarial examples for image segmentation, operating like a per-pixel targeted version of FGSM. DAG uses an incorrect segmentation map, an input image, and a target set of non-background pixels. The aim is to compute a minimum perturbation r that will change the pixel-wise prediction from the ground truth class to the incorrect target class.

DAG utilizes (1) an incorrect segmentation mask $\hat{y} = \{\hat{y_1}, \ldots, \hat{y_n}\}$ for an image $\boldsymbol{x}$ (2) the ground truth mask $y = \{y_1, \ldots, y_n\}$, where $y_n \in \{1, 2, \ldots, C\}$ and C is the number of classes and (3) a set of N pixels $T = \{t_1, t_2, \ldots, t_n\}$. In semantic segmentation T is composed of all pixels of the image but in order to constrain the search-space of the perturbations we limit T to the non-background pixels of the image.

The goal of DAG is to minimize the distance between the prediction of the ground truth and the incorrect target, as can be shown in [301]:

$$L(\boldsymbol{x}, T, y, \hat{y}) = \sum_{n=1}^{N} [z_{y_n}(\boldsymbol{x}, t_n) - z_{\hat{y}_n}(\boldsymbol{x}, t_n)], \tag{6.1}$$

where $Z = \{z_1, \ldots, z_C\}$ are the logits of the model, i.e. the output of the classifier f before the Softmax activation function. In step m the image has been perturbed to $\boldsymbol{x_m} = \boldsymbol{x} + \sum_{m=0}^{M} r_m$, where the perturbation r_m is computed by:

$$r_m = \sum_{t_n \in T} [\nabla_{\boldsymbol{x_m}} z_{\hat{y}_n}(\boldsymbol{x_m}, t_n) - \nabla_{\boldsymbol{x_m}} z_{y_n}(\boldsymbol{x_m}, t_n)]. \tag{6.2}$$

We use DAG to create adversarial images, as can be seen in Fig. 6.6, by selecting targets with varying degrees of difficulty. Notably, in Type A attack, we set the target to consist of background pixels; in Type B, we randomly assign 2000 pixels of the image to a randomly chosen adversarial class, and in Type C, we dilate a particular target class keeping all other classes unchanged. In our case, the class that was dilated was the skull. Of the described attack types, Type A is the most difficult, causing the highest amount of perturbation, while Type C is expected to cause the lowest distortion to the image, as demonstrated in Fig. 6.6.

In order to make sure that the adversarial perturbations were imperceptible, we measured the Mean Square Error (MSE) between the original and adversarial examples. The MSE remained extremely low, ranging from 0.004 for adversarial examples of Type A to 0.002 for Types B and C.

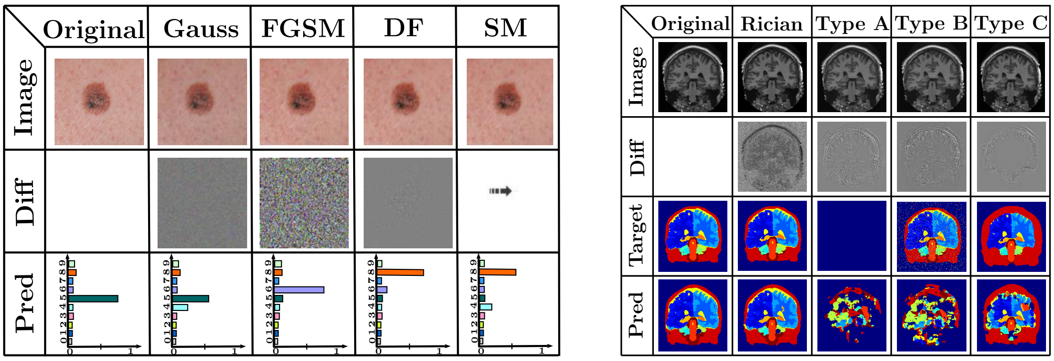
The introduced technique for evaluating model robustness includes benchmarking DNNs against task-specific adversarial attacks and is consistent across tasks. In the case of classification, we created adversarial examples with FGSM, DeepFool, and Jacobian Saliency Map Attack, while for segmentation, we used DAG with the three different types of targets (Type A-C) described above. We created adversarial examples only from the test set of each dataset, which was non-overlapping with the training set and only used the adversaries during inference time.

We selected a black-box threat model for the adversarial crafting, meaning that we first trained DNNs to craft the adversarial examples and then attacked independently trained DNNs that were not used for any adversarial crafting.

Contrasting with Noise

It could be argued that applying random noise on the test samples at inference time could replace adversarial examples. However, challenging ambiguous images and outliers cannot be modeled by noise distributions. Adversarial examples, which are created purposefully to cause failure in DNNs, are more suitable for investigating the behavior of a model when subject to extreme input changes.

To validate this statement, we also created samples perturbed with modality-specific noise to highlight further how adversarial perturbations differ from random noise. In the case of dermatoscopic images, we selected Gaussian noise and for T1w MRIs Rician noise. For a fair visual comparison, the Structural Similarity (SSIM) between the clean and noisy examples was the same as the one between the original and adversarial examples, ranging from 0.97 to 0.99.



(Left) (Right)

Fig. 6.6. Overview of adversarial examples along with their impact on the predictions of the model. Left: Fine-grained skin lesion classification and Right: Whole brain segmentation. The arrow for the SM attack indicates the minimal amount of pixels that needed to be perturbed in the original image. Unlike the predictions on original images, the adversarial examples successfully fool the models into either wrong classification or predicting incorrect segmentation masks. Figure published in [215], used with permission from Springer Nature Customer Service Centre GmbH.

To highlight the difference between clean and noisy images in the feature space, we show the t-Stochastic Neighbor Embedding representation (t-SNE) from InceptionV3 for clean, noisy, and adversarial examples crafted with FGSM in Fig. 6.7 for the classification task.

Contrasting Fig. 6.7(L) with Fig. 6.7(R), we can observe that images perturbed with noise are embedded close in space to clean samples, in contrast to adversarial examples that are pushed further towards other classes. The adversarial examples' behavior strongly supports that they do not resemble random noise and can serve as a strong benchmark for assessing a model's robustness.

6.3.3 Experiments

For the proposed robustness evaluation scheme, we performed fine-grained skin lesion classification and segmentation of the whole brain. The task-specific DNN training is described below:

Classification

For this task we fine-tune three commonly used deep learning architectures specifically, InceptionV3 [272], InceptionV4 [272] and MobileNet [129]. Both IInceptionV3 and InceptionV4 are particularly deep architectures ($>$ 100 layers), while MobileNet is significantly more compact, suitable for deploying on mobile devices as suggested by its name. We selected these models because comparing these architectures can uncover a relationship between model complexity, in terms of depth and number of parameters, and robustness.

For a fair comparison, all models were initialized with their respective ImageNet parameters and trained with a weighted cross-entropy loss with random data augmentation with affine transformations. The class weights were computed using Median Frequency Balancing [238] to combat the severe class imbalance. The models were optimized using stochastic gradient

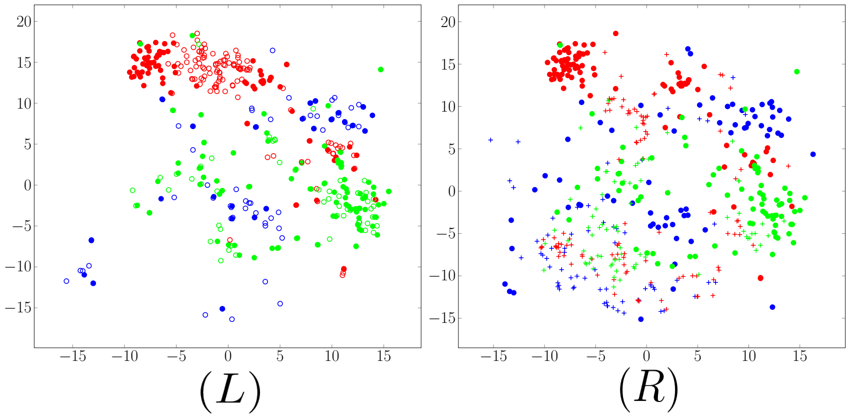


Fig. 6.7. t-SNE representation of the embeddings of IV3 for 3 classes of dermatoscopy images (shown in red, blue, and green) from clean (•), noisy (○) and adversarial images (+) crafted with FGSM. The noisy examples (○) are closer to the clean data in the embedding space (L), while adversarial samples are pushed to the boundaries and remain further away from the clean samples in the embedding space(R). Figure published in [215], used with permission from Springer Nature Customer Service Centre GmbH.

descent with a decaying learning rate initialized at 0.01, momentum of 0.9, and dropout of 0.8 to limit overfitting. We utilize the publicly-available Dermofit [19] image collection consisting of 1300 dermatoscopic images, with histologically validated expert annotations splitting them into 10 classes, namely Actinic Keratosis, Basal Cell Carcinoma, Melanocytic Nevus, Seborrhoeic Keratosis, Squamous Cell Carcinoma, Intraepithelial Carcinoma, Pyogenic Granuloma, Haemangioma, Dermatofibroma and Malignant Melanoma. The dataset was split patient-level with non-overlapping folds. 50% of the images were used for training, and the rest 50% for testing.

Segmentation

Regarding segmentation we evaluated three widely used fully-convolutional architectures, specifically SegNet [16], UNet [236] and DenseNet [136]. Among the selected models, we explore the impact of skip connections to robustness ranging from no skip connections at all in SegNet to long-range skips in UNet and both long and short-range skip connections in DenseNet. The network parameters concerning depth and layers were set to maintain comparable model complexity to investigate the impact of skip connections on model robustness.

The segmentation models were trained with a combined loss function, consisting of equally contributing weighted-cross entropy and Dice loss. The class weights in the cross-entropy loss were computed using Median Frequency Balancing following [238]. The models were optimized with Adam optimizer [154] with an initial learning rate of 0.001. We leverage 27 MRI T1 scans from the publicly available whole-brain segmentation Multi-Atlas Labeling Challenge (MALC) Dataset [186]. MALC is a subset of Open Access Series of the Imaging Studies (OASIS) dataset [187] that was released in MICCAI 2012 [166]. We use an 80-20 patient-level split for training and testing. We segment the brain scans to 15 different structures with manual expert annotations provided by Neuromorphometrics, Inc. with an academic subscription.

The models for both tasks were trained using TensorFlow [2], and adversarial examples for DeepFool and Saliency Map Attacks were creating using the FoolBox [232] library.

The model performance and robustness are evaluated using classification accuracy for the skin lesion classification and Dice score for the brain segmentation. Dice was calculated volume-wise, without considering the background class, and averaged across test subject volumes. The overall performance of the models on clean and noisy examples is reported in Table 6.1. The average score against all adversarial attacks is also shown for easier comparison and the drop in performance after the attacks. Moreover, Table 6.2 and Table 6.3 show the performance of each model for all individual black-box attacks.

6.3.4 Results and Discussion

Robustness Evaluation for Classification

Below we will discuss the results of the evaluation on the classification task.

Visual Evaluation

Fig. 6.6 (Left) shows adversarial images of an unseen test example of the class malignant melanoma for each classification attack, namely FGSM, DeepFool, and Saliency Map Attack, along with an image distorted with Gaussian noise for comparison. A scaled version of the difference between the original and adversarial image is illustrated along with the class probabilities of InceptionV3.

Regardless of the attack, all adversarial examples are consistently assigned to the wrong class with high confidence by the model. However, adding Gaussian noise results only in confidence reduction, while the model prediction remains correct. Moreover, it is shown that FGSM creates perturbations across all pixels on the whole image, while DeepFool and Saliency Map Attack craft perturbations localized to the lesion area. Specifically, the Saliency Map Attack only perturbs very few pixels of the whole image while still fooling the target model.

Attacks

Table 6.1 shows that InceptionV4 and MobileNet both reach comparable accuracy on the clean data (0.81/0.80), which is higher than InceptionV3 (0.71). Considering only a model's performance on clean data, one could conclude that IV3 achieves the worst comparative performance. However, when comparing the robustness of these models regarding their average performance under all attacks, we observe a different trend. The average performance drop across all the attacks for InceptionV3 is substantially lower (-6.89%) in comparison to InceptionV4 (-17.72%) and MobileNet (-24.55%).

Comparing each attack individually in Table 6.2 we can see that Inception V3 outperforms both the other models for FGSM by 0.1 compared to InceptionV4 and 0.3 against MobileNet. For DeepFool, InceptionV4 achieved the highest robustness by a small margin of 0.05. Regarding Saliency Map Attacks, InceptionV3 and InceptionV4 performed similarly and had a marginally lower drop in performance than MobileNet.

Tab. 6.1. Comparative evaluation of the classification and segmentation models on clean, noisy and adversarial examples. We report the average accuracy for the classification task and Dice Score for segmentation along with the % drop in performance on adversarial examples with respect to performance on the clean test set. Table published in [215], used with permission from Springer Nature Customer Service Centre GmbH.

	Classification			
	Clean	Gaussian Noise	Adversarial Average	% Adversarial Drop
InceptionV3	0.710	0.693	**0.641**	**-6.897**
InceptionV4	**0.810**	**0.761**	0.633	-17.72
MobileNet	0.800	0.647	0.564	-24.55
	Segmentation			
	Clean	Rician Noise	Adversarial Average	% Adversarial Drop
SegNet	0.842	0.595	0.470	-37.17
UNet	**0.862**	0.759	0.453	-40.92
DenseNet	0.861	**0.848**	**0.667**	**-19.53**

Contrasting InceptionV4 and MobileNet on their performance on the adversarial attacks, it is clear that InceptionV4 shows enhanced robustness capabilities, even though their performance on the clean dataset differed by only 0.01. Furthermore, even though InceptionV3 had 0.1 lower accuracy than InceptionV4 on the clean data, it achieved comparable robustness and even outperformed InceptionV4 for FGSM attacks.

Performance on Random Noise

Table 6.1 shows that the drop in performance caused by the Gaussian noise distortion was substantially lower in comparison to the adversarial perturbations. The accuracy of InceptionV4 and InceptionV3 only dropped by 0.05 and 0.01, respectively, showing the robustness of these models to noise. MobileNet's accuracy dropped by 0.15, showing the decreased capabilities of this model compared to the other two, both in terms of random and adversarial perturbations.

Architecture Comparison

From our experiments, we found that InceptionV3 and InceptionV4 achieved higher robustness than MobileNet. That could be attributed to the fact that MobileNet's shallow architecture could not learn the underlying distribution of the training data properly. InceptionV3 and V4 were comparable in terms of robustness to noise and attacks; however, InceptionV4 achieved overall better performance in the cleat test set. As we discussed above, a combination of a deeper architecture like InceptionV3/4 with a pruning or quantization technique [109, 176] could lead to a good overall solution for models that perform successfully both on clean and adversarial inputs.

Tab. 6.2. Comparative evaluation of model robustness using black-box attacks for the task of **classification**. The average accuracy is reported. Table published in [215], used with permission from Springer Nature Customer Service Centre GmbH.

	FGSM			**DeepFool**			**SMA**		
	IV3	IV4	MN	IV3	IV4	MN	IV3	IV4	MN
InceptionV3	**0.449**	**0.548**	**0.567**	0.729	0.707	0.664	**0.738**	0.701	0.669
InceptionV4	0.429	0.411	0.451	**0.743**	**0.768**	**0.697**	0.735	**0.778**	**0.683**
MobileNet	0.335	0.275	0.213	0.726	0.731	0.672	0.732	0.735	0.661

Tab. 6.3. Comparative evaluation of model robustness using black-box attacks for the task of **segmentation**. The average Dice score is reported. Table published in [215], used with permission from Springer Nature Customer Service Centre GmbH.

	Type A			**Type B**			**Type C**		
	SegNet	UNet	DenseNet	SegNet	UNet	DenseNet	SegNet	UNet	DenseNet
SegNet	0.277	0.272	0.309	0.397	0.473	0.428	0.669	0.702	0.705
UNet	0.248	0.434	0.258	0.364	0.434	0.368	0.636	0.653	0.677
DenseNet	**0.600**	**0.528**	**0.415**	**0.749**	**0.721**	**0.563**	**0.819**	**0.791**	**0.814**

Furthermore, Table 6.2 showed that all attacks were successful to all models regardless of the architecture that was used to craft them. This showcases the transferability of adversarial examples and the close decision boundaries of different DNNs [282].

Robustness Evaluation for Segmentation

Below we will discuss the performance of the selected segmentation networks in terms of generalizability and robustness.

Visual Evaluation

Fig. 6.6 (Right) illustrates how the prediction maps of the trained DN model change with adversarial input. Initially, we can see that the model's prediction on the original input is very close to the ground truth. However, the crafted DAG attacks of Type A-C successfully influence the model into predicting incorrect segmentation maps.

Specifically, Type A attack is severely influencing the model and has eliminated even the general structure of the brain from the prediction. Type B similarly distorted the prediction and, as expected by the shuffled pixels on the target, the model predicts pixels from every class scattered across the segmentation map. Finally, Type C mainly affected the skull segmentation but also decreased the quality of the neighboring structures.

The prediction on the image distorted with Rician noise is visually similar to the one of the original image and the ground truth. This demonstrates that adding adversarial perturbation on an image does not resemble distorting it with random noise.

Attacks

In Table 6.1, we see that DenseNet (0.861) and UNet (0.862) achieve nearly equal performance on the clean unseen test set and perform better than SegNet (0.842). However, the performance gap increases dramatically for the adversarial attacks. DenseNet outperforms both UNet and SegNet by 0.2 and achieves the lowest drop in Dice score of -19.53%. UNet and SegNet achieve similar performance under attack, with a drop of -37.17% for SegNet and -40.92% for UNet. We notice that even though UNet had the highest Dice Score on the clean test set, it has the lowest one under attack, showing its vulnerability.

In Table 6.3 we can see the superiority of DenseNet in comparison to the other models for all attack types. As expected, Type A causes the highest drop in Dice score, and the predictions of SegNet and UNet have a Dice of around 0.2 to 0.4. However, DenseNet manages to maintain a higher Dice of 0.4 to 0.6. The Type B attack drops the performance of DenseNet by 0.1 to 0.2, while the drop for Unit and SegNet is comparable and ranges between 0.4 to 0.5. Type C is the least powerful attack, causing the smallest drop for all models. However, DenseNet remains more robust than UNet and SegNet.

Moreover, the fact that the performance drop caused by the addition of Rician noise remains low both for UNet and DenseNet (10% and 1% respectively) reinforces the contrast between noise and adversarial perturbations.

In Fig. 6.8 we visualize the predictions of UNet and DenseNet for a Type C attack for three-volume slices. DenseNet predicts dilation in the skull for all three slices, as expected by the attack target. However, the regions inside the brain remain, for the most part, correctly segmented in comparison to the ground truth. However, UNet's predictions have lower quality, distorting the classes within the brain and the skull. These findings showcase the superiority of DenseNet in terms of robustness, even though both models achieve the same performance on clean data.

Performance on Rician Noise

In agreement with the classification results, the performance of all segmentation models to Rician noise, shown in Table 6.1 was higher than their performance on adversarial attacks. Specifically, DenseNet had the highest Dice score of 0.848, which was only 0.02 lower than its Dice on clean samples. UNet had a 0.11 Dise score decrease for Rician noise, while SegNet's Dice score dropped by 0.25. This experiment further underlined the superior robustness of DenseNet in comparison to UNet, even though UNet had a higher Dice score by 0.01 for clean data.

Architecture Comparison

Across the board, UNet and DenseNet outperformed SegNet for clean, noisy, and adversarial samples. This showcases the improvement brought upon the models using skip connections, which offer increased trainability and better gradient flow [117]. Based on the experiments on noisy and adversarial data, DenseNet outperformed UNet, showing that dense skip connections

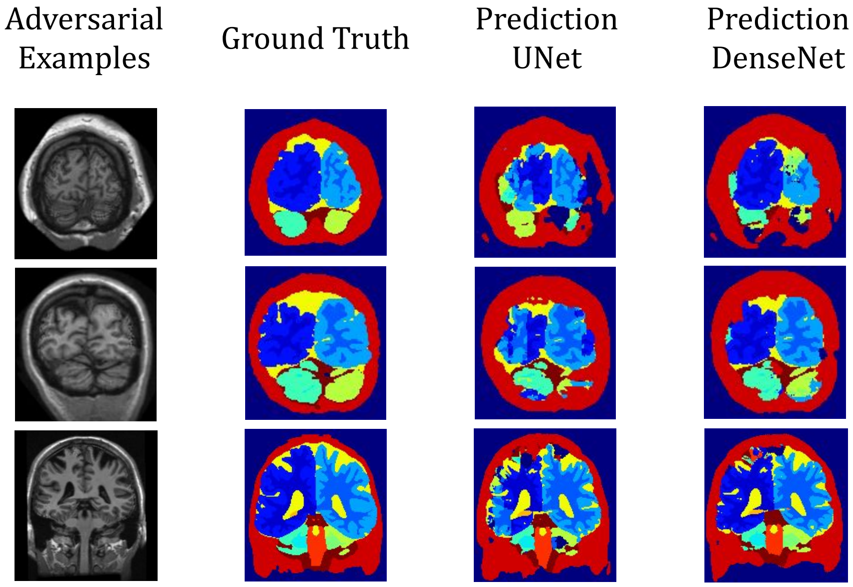


Fig. 6.8. Qualitative assessment of the predicted segmentation maps of UNet and DenseNet under attack Type C - skull dilation. Even though UNet achieves the same Dice score on the clean test set, its predictions are substantially more distorted in comparison to DenseNet.

further increase the robustness of the model and contribute to a better understanding of the underlying data distribution.

Moreover, the transferability of adversarial attacks across models is also shown in Table 6.3, where attacks of all source models cause comparable drop inaccuracy to all target models.

Overall our experiments showed that comparing the three segmentation models based on their generalizability would not have been adequate to determine the best model. Only after benchmarking the models with adversarial attacks and noise were we able to determine the best one. Both its resilience to samples distorted with Rician noise and its consistent resistance to adversarial attacks make DenseNet the most robust model among its competitors for this task.

Part IV

Robustness Beyond Imaging Data

Learning Beyond Imaging Data 7

7.1 Tabular Data for Medical Diagnosis

So far in this dissertation, we have discussed contributions and related work in the field of medical imaging. However, various medical data, such as electronic health records, written assessments, metadata regarding a patient's age, sex, or more, are tabular information. Methods that can successfully perform a diagnosis or treatment planning based solely on tabular data or combining tabular and imaging data could benefit the healthcare system and mine the patient information effectively. Furthermore, it has been shown that providing additional information to a DNN along with images in the form of the volumes of each brain region [174] contributes to the improvement of robustness.

Moreover, medical information is, in many cases, longitudinal since it is common to monitor the progress of a disease or the development of a child. Combining such longitudinal data for every patient would increase a DNN's understanding of the training distribution and allow for enhanced performance and robustness.

Thorough evaluation of a DNN can be achieved with model benchmarking, as discussed earlier. However, statistically analyzing the results of a model and interpreting its decisions can also contribute towards ensuring that the model is making correct predictions based on meaningful input features.

To that end, in this Chapter, we will discuss the final contribution of this dissertation. We perform depression score prediction for adolescents using solely longitudinal tabular data. Afterwards, we thoroughly evaluate the model's predictions and statistically analyze its performance. Our novel pipeline provides not only a DNN prediction score but an overall analysis of the model's behavior.

7.2 Related Work

7.2.1 Learning from Metadata

Medical diagnosis relies heavily on combining information from multiple sources; these include imaging data, laboratory data, or observational data. Medical image interpretation provides important clinical context that is often essential for diagnosis [170]. However, it has been shown that lack of clinical and laboratory during image interpretation can lead to lower radiologist performance [32].

Data fusion describes the combination of data from multiple modalities in order to extract complete information that could increase the performance and robustness of machine learning models in comparison to using a single modality. The most commonly used strategies for combining multi-modal information, such as imaging and tabular data, are early, joint, and late fusion and can be seen in Fig. 7.1

Early fusion includes joining multiple inputs into a single feature vector before feeding into a single model for training. The aggregation can be performed using concatenation or pooling. The aggregated features could be used without any pre-processing (Type I) or could first be pre-processed by a manual or learned feature extractor (Type II) [131].

Joint fusion describes the combination of learned feature representations from intermediate layers of DNNs with features from other input modalities. Feature extraction can be performed for one of the input modalities (Type II) or for both inputs (Type I). The main difference with early fusion is that the loss of the final Models is propagated to the feature extraction CNNs, enhancing the extracted feature quality during training [131].

In Late Fusion [230] different input modalities are used to train separate DNNs, and then the final prediction is made with an aggregation function that combines the predictions of multiple models. The aggregation functions can be averaging, majority voting, weighted voting, or a meta-classifier.

7.2.2 Longitudinal Predictions

Longitudinal data has been used in various medical imaging applications to predict the progression of Alzheimer's disease [6, 27], for survival analysis [168] and for lung infection progression [152].

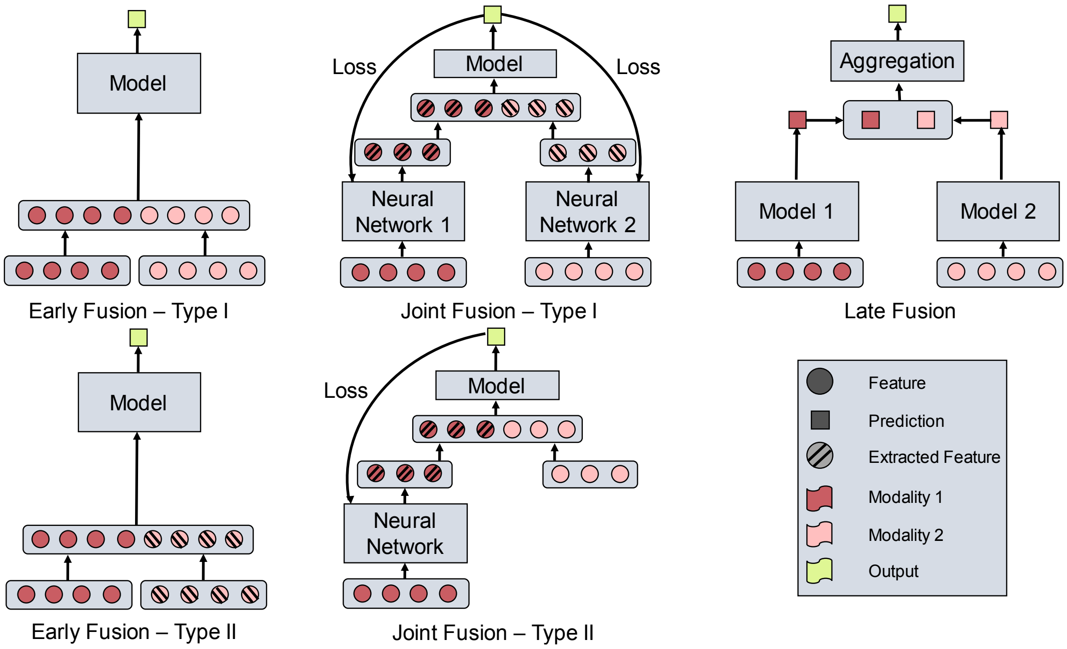


Fig. 7.1. Overview of different types of multi-modal fusion [131]. Early fusion joins original or extracted features at the input level. Joint fusion combines features at the input, but the loss is propagated back to the feature extraction model. Late fusion joins the predictions at the decision stage.

Aghili et al. [6] utilized recurrent neural networks (RNNs) to analyze regression patterns from longitudinal data with missing variables and perform classification between Alzheimer's Disease, Mild Cognitive Impairment, and healthy subjects. For this task, they successfully used two variations of RNNs, namely Long Short Term Memory and Gated Recurrent Unit.

In our work with Kim et al. [152] we analyzed longitudinal CT Scans from patients suffering from COVID-19. A longitudinal segmentation network was used to identify the regions of healthy lung, consolidation, GGO, and pleural effusion. Experiments showed that analyzing longitudinal information significantly improved the performance of a DNN in comparison to the use of static data.

7.2.3 Adversarial Attacks on Tabular Data

Adversarial attack crafting has not been as explored for the field of tabular data as for the imaging inputs. There have been few attacks proposed on tabular data for fraud detection that will be discussed below.

Cartella et al. [52] proposed a model agnostic attack applicable to any architecture, even decision trees for fraud detection systems. They also discuss the challenges of crafting a tabular data attack. Specifically, image data normally vary within a limited range and data types; however, tabular data include different types of information, such as demographic information, surnames, or amounts. Even if all values are numerically encoded, it is still challenging to create realistic and imperceptible adversarial attacks for tabular data. Their attack was based on the Zeroth Order Optimization (ZOO) [56] algorithm modified for tabular data with class imbalance. Their attacks were able to decrease the performance of fraud detection systems dramatically.

Ballet et al. [20] focused on the imperceptibility of the tabular adversarial attacks. Their method, called LowProFool (low profile) Attack finds the features of the tabular data with the highest importance using Pearson's correlation coefficient and adopts a black-box gradient-based attack method applicable to DNNs.

PermuteAttack [113] introduced a counterfactual example generation method for tabular data with discrete and categorical variables. The proposed algorithm used a gradient-free optimization based on genetic algorithms and was applicable to any classification model.

Finally, An et al. [8] introduced Longitudinal AdVersarial Attack (LAVA) on electronic health records tabular data. LAVA introduced a small amount of perturbations on clinical features that were not likely to be detected based on the Jacobian Saliency Map Attack [213] and a dual attention mechanism. Their attack had a low detectability rate as the perturbations were spread across multiple visits and features of each subject.

Overall, tabular data include nominal, ordinal, and real-valued data. Furthermore, different features have substantially different value ranges. Oftentimes, tabular data have missing values, and the interactions between features are complex [171]. All those factors increase the difficulty of designing successful adversarial attacks on tabular data. Therefore it is beneficial to include such data in the input of a DNN to increase the difficulty of an attack and enhance the model robustness [174].

To this end, in this work, we focus on analyzing tabular, longitudinal data using recurrent neural networks.

7.3 Depression Score Prediction from Longitudinal Tabular Data

7.3.1 Introduction

During adolescence, the prevalence of major depressive disorder (MDD) increases from 1–3% to 20% by the age of 18 [15]. Heightened vulnerability and early onset of mood disorders during adolescence are hypothesized to relate to the dramatic developmental changes that occur during these sensitive years [169]. This period is characterized by key behavioral, hormonal, and physical changes linked to puberty, brain plasticity [93, 97, 270], changes in circadian and homeostatic bioregulatory processes [309], and increased susceptibility to psychosocial stressors.

Depression during adolescence is associated with long-term clinical course [89], anxiety disorders [196], sleep disturbances [253], eating disorders [127], substance use [149], and suicide attempts, with trajectories extending into adulthood [96]. Given the prevalence and considerable costs of depression in this population, there is an urgent need to identify the risk and protective factors of depression in youth, considering the presence of depressive symptoms that may preclude the development of major depressive disorder. Because of the variation of symptoms, clinical subtypes, and the prevalence of subclinical and mild depression [252], it has been suggested to shift the focus of research on the continuums on which symptoms occur, rather than the diagnostic criteria [133].

The RDoC (NIMH Research Domain Criteria - RDoC) [134] framework provides a transdiagnostic approach to symptoms common across several disorders, integrating recent advances in genetics, neuroscience, and cognitive science. This approach introduces six systems encompassing broad domains of human functioning, namely the Positive Valence Systems, Negative Valence Systems, Cognitive Systems, Systems for Social Processes, Arousal/Regulatory Systems, and Sensorimotor Systems. Inspired by the RDoC matrix, the current work focuses on two topics central to depressive disorders—namely, **anhedonia** (as part of the Positive Valence Systems domain of reward learning) and **negative valence** (covering areas of the Negative Valence Systems domain).

Given the rapid transitions from dependence on parents to relative independence in the developmental context of middle and late adolescence, understanding the effects of different psychosocial and cognitive factors on the risk for youth depression could be particularly important to advance effective preventive and intervention efforts. There is growing evidence suggesting that in addition to the developmental changes, personality traits [157], stressful life events [193], social relationships [91, 278], sleep health [180, 253], and cognition [191, 222] also play a critical role in the onset and maintenance of depressive disorders. Thus, elucidating developmental risk models for depressive symptoms requires investigation of multiple psychosocial and behavioral constructs that interact to increase the risk of mental health problems.

The psychiatric symptom of anhedonia reduced pleasure, and interest in previously enjoyable, rewarding experiences, is a key symptom of major depressive disorder [25, 115]. Persistent anhedonia in childhood and adolescence is an important predictor of adult-onset MDD [293]. It is considered a motivational, reward-processing deficit which might be the result of underlying brain level reward system dysfunctions [283].

Furthermore, depression is often associated with feelings of sadness and loss, which can be responses to frustrating and unpleasant situations, such as sustained anxiety, fear, threat. These feelings and responses can be combined under the broad construct of negative valence [71]. Symptoms of negative valence are less responsive to antidepressants [79, 194] than anhedonia, and are related to different clinical features [194]. Moreover, depressed individuals tend to a negativity bias in information processing, which is reflected in lower valence ratings for emotional faces, especially for happy and neutral faces [73].

In this work, we used a framework based on a recurrent neural network to track anhedonia, and negative valence across adolescence in a large sample of participating in the National Consortium of Alcohol and Neurodevelopment in Adolescence (NCANDA) [38], to determine the predictive value of different psychological and life domains. Importantly, and following the RDoC approach, the current model integrates the psychosocial information about individuals over time and takes the mood history of the person into account when making predictions [244]. We report findings based on two of the RDoC units of analysis, namely, behavior and self-reports. The current longitudinal approach helps map developmental trajectories that differentially lead to psychological dysfunction reflected in anhedonia and negative valence. Furthermore, along with the dysfunctions we predict a subject's age in a multi-task fashion, which allows the proposed model to take the developmental context into account during the classification task.

Contributions:

- We utilize a longitudinal multi-task model and accurately predict depression scores in adolescents.
- We leverage permutation testing and identify the most significant input measurement categories.
- We use model interpretation and pinpoint the individual input variables that contributed the most to the model's decision.
- We perform statistical analysis and validate the significance of the most important input variables for the model.

7.3.2 Predicted Measurements

Anhedonia

At each assessment, psychiatric symptoms were evaluated in participants using the Achenbach System of Empirically Based Assessments (ASEBA; [4]). As a measure for the construct, within the Positive Valence Systems [23], we used a single item for anhedonia: "There is very little that I enjoy," dichotomizing the items into 0 (not true) and 1 (sometimes or often true), as in [25].

Negative Valence

As a composite measure of acute threat (fear), potential threat (anxiety), and loss constructs within the Negative Valence Systems, we used the depression/anxious subscale, which is comprised of 13 items, including being fearful/anxious, nervous/tense, and cries a lot. Normalized T-scores were calculated based on age and sex, and a dichotomized variable was created ($>$ T-score of 65 = depressed/anxious).

7.3.3 Input Measurement Categories

As input to our recurrent neural network, for every subject we used a set of measurements in the form of tabular data that were acquired multiple times over the course of their adolescence. The measurements can be grouped in the following categories:

Sleep

Sleep characteristics [114] were assessed through several measures taken from established measures [200, 259], assessing morningness, eveningness [43], sleep timing and duration [200], and finally, sleep disturbance [43]. Circadian preference was assessed using an abbreviated 4-item version (CSM-4) of the Composite Scale of Morningness (CSM; [259]). The CSM-4 score ranges from 4 to 18, with higher scores indicating greater morningness. Sleep disturbance was assessed using a single item ("During the past month, how would you rate your sleep quality overall?") drawn from the Pittsburgh Sleep Quality Index [43]. Response options ranged from 1 to 4 in this order: "very good," "fairly good," "fairly bad," and "very bad." Finally, habitual sleep timing (bedtime and rise time) and sleep duration were assessed separately for weekdays and weekends, and weekday-weekend shifts in sleep timing (weekend minus weekday) were calculated separately for bedtime and rise time.

Personality

Personality and temperament were assessed using the UPPS-P Impulsive Behavior Scale (UPPS-P [72]) and the Ten-Item Personality Inventory (TIPI [105]). The UPPS uses 20 statements (scale of 1-4) to examine facets of impulsivity, including behaviors influenced by changes in effect (e.g., positive and negative urgency), lack of planfulness, and behavioral persistence. The TIPI assesses broad personality domains of Conscientiousness, Agreeableness, Extraversion, Emotional Stability, and Openness to Experiences.

Moreover, cognitive coping strategies employed to manage interpersonal stressors were measured using the Response to Stress Questionnaire (RSQ) [68]. Participants rated how often they used each coping method or experienced each type of involuntary stress response on a scale of 1 (Not at all) to 4 (A lot). The RSQ contains items to measure three types of coping mechanisms and two types of involuntary stress responses.

Substance Use

Subjects participated in the Customary Drinking and Substance Use Record (CDDR [40]) to characterize their alcohol and substance use history and current use. This measure includes questions about use frequency, and the maximum number of drinks in a drinking episode, during the past year. Using these data, NCANDA participants were defined at study entry as "no/low" drinking (majority of the sample, 83%) or "exceeds criteria" (17%). Past year alcohol use data were also used to categorize participants as heavy, moderate, and no/low drinkers using the modified [44] inventory, comprising quantity (average and maximum consumption) and frequency combinations.

Participants also completed a modified Semi-Structured Assessment of the Genetics of Alcoholism (SSAGA; [42, 124]), which assessed Axis I diagnoses, including alcohol and substance use, mood, anxiety, and conduct, disorders. Participants and one parent (for participants under age 18) completed the self-report instruments from the Achenbach System of Empirically Based Assessments (ASEBA). Participants who endorsed one or more symptoms of conduct disorder or antisocial personality disorder were considered high risk (externalizing symptoms) for alcohol use and problems (21%).

Internalizing symptoms were assessed with the SSAGA and participants who met criteria (28% of the total sample) were considered at risk for alcohol use and problems (internalizing symptoms).

Furthermore, family history of substance use disorders was assessed with the Family History Assessment Module (FHAM; [234]). Family history positive participants had at least one biological parent with problems related to alcohol/drug use; or two or more biological grandparents with significant problems related to alcohol/drug use; or one or more biological grandparent and two or more other biological 2nd-degree relatives with significant problems related to alcohol/drug use. Using these criteria, 17% of the sample was positive for familial alcohol use problems and 8% for familial drug use problems [38].

Life

The Childhood Trauma Questionnaire (CTQ; [26]), a reliable and validated measure of self-reported traumatic experiences during childhood, was used. It consists of five subscales about negative childhood experiences, each comprised of 5 items: emotional and physical neglect and physical, emotional, and sexual abuse. Higher scores indicate increased severity of childhood maltreatment.

We used the Adverse Childhood Events - International Questionnaire [5], a 29-item measure that assesses exposure to three domains of childhood adversities: childhood maltreatment, family/ household dysfunction, and violence outside the home. Respondents are asked to respond to questions about their experiences during the first 18 years of their lives.

Support

The self-report California Healthy Kids Survey (CHKS) [46] was utilized to assess the health risk behaviors and resilience information. The overall instrument contains 65 items, with responses on a 4-point Likert scale, ranging from “not at all true” to “very much true.” We used the following items: “Who really cares about me.” (adults at school), “Who always wants me to do my best.”, “Whom I trust.”, “I am part of clubs, sports teams, church/temple, or other group activities.”, “I am involved in music, art, literature, sports, or a hobby.”, “I help other people.”

Executive dysfunction - BRIEF

To assess the inhibitory control and flexibility of adolescents, we utilized the self-reported variant of the Behavior Rating Inventory of Executive Function (BRIEF-SR: [98]). BRIEF-SR comprises 80 items distributed into eight subscales designed to measure impulsivity across the domains of executive functioning.

Neuropsychological measures

In the Delay Discounting task [263] participants made preference judgments about accepting a small immediate reward ($100.00) over a larger hypothetical ($1000.00) reward given at varying delays (e.g., 1 day, 1 week, 1 month, or 6 months) in the future. The immediate reward is automatically adjusted according to the participant's choices; it increases if the future reward was chosen or decreases if the immediate reward was just chosen. Impulsivity or impulsive decision-making is associated with a higher tendency to devalue larger longer-term gains over short-term benefits, and it is reflected by a steeper discounting rate. Further details of the test administration are described in [268].

The Stroop task, originally proposed by Stroop [266] measures the ability to inhibit automated reading responses. The Stroop Match to-Sample task [246] was administered and consisted of two conditions. The task required adolescents to match the color of a sample stimulus to the color of a Stroop word. The participants were required to read names of colors and respond with "Yes" if the displayed word's font color matches the initially presented sample color and respond with "No" when the sample and the target word's font color do not match.

Attention and working memory were measured by the Short fractal N-back test [155]. In the 0-back condition, the target was a fractal design displayed on the computer screen that matched a pre-specified fractal image. This condition requires sustained attention with low working memory load.

The Penn Emotion Recognition Test (PERT: [110]) measures the ability to identify six basic emotions in facial expressions displayed on a screen. The computer-morphed target images (40-item) that are derived from the facial features of real individuals (20 female and 20 male), each showing a specific emotion. The images present angry, scared, happy, sad, and neutral faces varying between low and high intensity. The task required participants to indicate the expressed emotion from a list of 5 choices.

7.3.4 Participants

The study consisted of 621 youths (ages 12 to 17 years at baseline) who were recruited by NCANDA from November 2012 to October 2014 across five sites: University of California at San Diego (UCSD), SRI International, Duke University Medical Center, University of Pittsburgh (UPMC), and Oregon Health & Science University (OHSU) [39]. The participants and parents provided written informed consent before participation in the study. The Institutional Review Boards (IRB) of each site approved data collection and use. Each subject participated in up to 7 assessments, with the average time between assessments being 1.05 years. It should be noted that, assessments were excluded from the analysis once the subject turned 18 years old.

Tab. 7.1. Demographics of the NCANDA cohort. ± denotes the average and standard deviation

General	
Sex (Female/Male)	310 / 311
Number of Assessments	3.20 ± 1.66
Time Between Assessments in Years	1.05 ± 0.15
Baseline	
Age in Years	15.02 ± 1.69
Pubertal Development Score (PBS)	3.04 ± 0.69
Body Mass Index (BMI; z-score)	0.32 ± 1.01
Parents Education in Years	16.88 ± 2.46
Race	
Caucasian	438 (70.53%)
Hispanic	74 (11.82%)
African-American	81 (13.05%)
Asian	38 (6.12%)
Other	64 (10.31%)
Site	
UCSD	154 (24.80%)
SRI International	146 (23.51%)
Duke	137 (22.06%)
OHSU	108 (17.39%)
UPMC	76 (12.24%)

At each assessment, participants completed a battery of neuropsychological and clinical assessments, which covered eight categories that were described above: personality, sleep, life, Behavior Rating Inventory of Executive Function (BRIEF) [99], neuropsych, substance use, support, and additionally, demographics. Demographics incorporates all variables listed in Table 7.1 except age.

Assessments of participants were divided into groups based on their self-reported emotion measures. Anhedonia was observed if the individual reported very little joy within the last 6 months of an assessment. The criteria for negative valence was whether the Anxiety/Depression trait T-score was 65 or higher or they often experienced unhappiness, sadness, or depression in the last six months.

Among the 621 youths, 116 individuals reported anhedonia, and 81 reported negative valence in at least one of their assessments. 51 youths reported at least once both constructs and 475 participants were viewed as controls as they reported neither construct

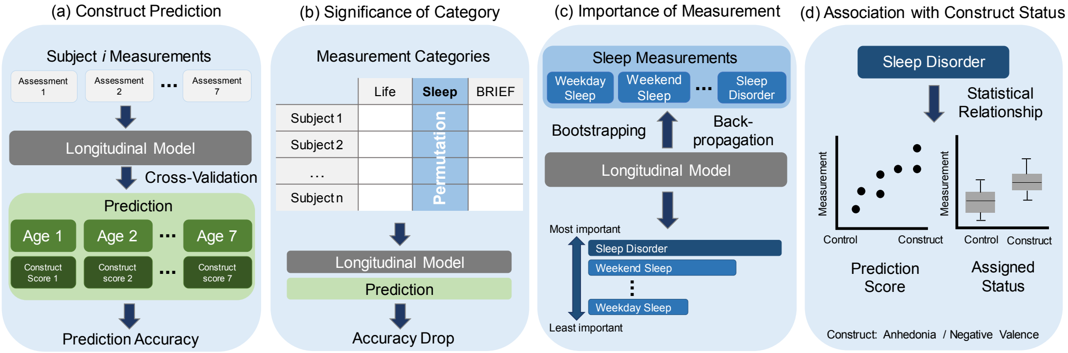


Fig. 7.2. Overview of the proposed pipeline. Annual assessments of 621 NCANDA youths split into control/anhedonia/negative valence groups were processed by a longitudinal neural network, which predicted each individual's age and diagnostic status at every assessment. Afterwards, the measurements of each variable category were permuted and reassessed by the trained model to identify the significance of each category. Furthermore, the individual measurements of the significant categories were ranked using the gradient magnitudes of the trained model. Finally, the measurements identified as most important, and the respective predicted scores and assigned status (control/anhedonia/negative valence) were correlated.

in any of the assessments. The analysis was performed based on the public data release NCANDA_PUBLIC_6Y_REDCAP_V01 [140].

7.3.5 Machine Learning and Statistical Analysis

First it should be noted that missing measurements in the neuropsychological and clinical assessments of the subjects were replaced with those of the nearest assessment for that individual. If the measurement was never recorded for the subject, the mean across all subjects was used to fill out the missing value.

Step 1: Anhedonia/Negative Valence Prediction

Separately for anhedonia and negative valence, the completed measurements were analyzed by a longitudinal deep learning model (Fig. 7.2(a)) consisting of a Fully Connected Layer [104] and a Recurrent Neural Network with a gating mechanism [62]. This model was trained and tested in determining the *age* and the confidence *score* (between 0 and 1) regarding the presence of a construct at an assessment of a subject via 5-fold stratified cross validation [11], i.e., dividing the subjects into 5 folds, selecting 4 folds for training the model and 1 fold for testing the model, and repeating training and testing until each fold was used for testing.

To evaluate the performance of the model Balanced Accuracy (BACC) was used since the dataset was characterized by class imbalance [37]. Specifically, the accuracy for each class was measured and then averaged across the two classes. The significance of the BACC (p-value $<$ 0.001) was computed using the Fisher exact test [88].

The model was trained on both age and anhedonia/negative valence score in a multi-task manner so that the ages of subjects across assessments could be implicitly aligned. As can be seen in Table 7.2, subject ages varied from 12 to 17 years old at every assessment. Furthermore, Tab. 7.2. Age of subjects at each assessment, starting from the baseline up to the 6-year follow-up under the age of 18. Notice the varying age at the baseline assessment, showcasing the need to provide additional age information to the model to implicitly align the visits. The number of annual assessment decreases from 621 for the baseline to only 4 subjects with 6 follow-up assessments.

	12	13	14	15	16	17	#Subjects
Baseline	92	111	104	109	108	97	621
Follow-up 1yr	2	86	99	103	103	96	489
Follow-up 2yr	0	4	77	101	91	101	374
Follow-up 3yr	0	0	3	76	90	98	267
Follow-up 4yr	0	0	0	1	70	90	161
Follow-up 5yr	0	0	0	0	1	67	68
Follow-up 6yr	0	0	0	0	0	4	4

there were more than 100 measurements used as input to the model. Thus, separating age from the input and modelling it as an auxiliary task would be beneficial for the longitudinal prediction.

Step 2: Category Permutation Testing

For each of the 8 aforementioned categories (personality, sleep, life, BRIEF, neuropsych, substance use, support, and demographics), its importance in the process of predicting a construct was determined via permutation testing [101] (Fig. 7.2(b)). Permutation testing randomly rearranged the values of each measurement only in that category among subjects in the test set. Afterwards the BACC of the trained model on the permutated data was recorded. It should be noted that all measurements from the 8 categories were forwarded to the model. However, only the measurements from one category at a time were permutated. This procedure was repeated 500 times to compute the percentage of trials (p-value) that resulted in BACCs at least as high as the original (unpermutated) accuracy. The impact of the category on the prediction process was then viewed as significant if the p-value was smaller than 0.05 (or less than 25 permutations with at least as high accuracy scores). With this evaluation step, the measurement categories that were most important were identified.

Step 3: Individual Feature Importance & Model Interpretation

In this step, the most important individual measurements for the model were identified. For each category that met the significance level for both anhedonia and negative valence, the influences of individual measurements of that category on predicting a construct were determined by performing 100 runs of bootstrapping [81] (Fig. 7.2(c)). Each run consisted of randomly selecting (with replacement) 475 controls and 116 subjects with anhedonia (or 81 with negative valence) and then training the prediction model on the resulting data set.

The importance of a measurement towards correct predictions was then quantified by its *magnitude* according to guided backpropagation [262]. Guided Backpropagation is usually used to interpret the model decision on imaging data, but it can also be applied in non-imaging,

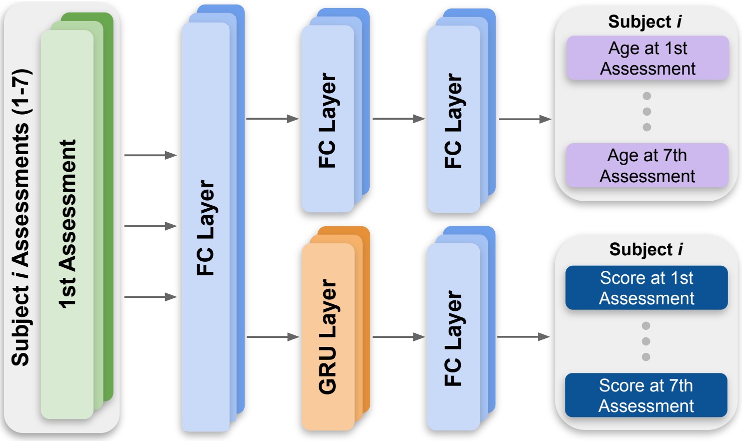


Fig. 7.3. Overview of the multi-task longitudinal model for depression score prediction from neuropsychological and clinical data for adolescents. All reports of one subject were jointly forwarded to the model. One branch consisted of a GRU [62] and a FC Layer [104] to perform longitudinal prediction of the status of anhedonia or negative valence. The other branch consisted of two FC layers and regressed the age of each subject at every assessment to encourage the alignment of subject ages across annual reports.

tabular data. Specifically, the magnitude of the model gradient with respect to the input features using backpropagation was calculated after setting all negative gradients to 0. After completing the 100 runs, each measurement within the significant categories was ranked according to their averaged magnitude across those runs.

Step 4: Correlation of Top Measurements with Predictions & Ground Truth

In this step, the aim was to validate, whether the measurements that were ranked with highest importance for the model, had meaningful correlation with the ground truth. To validate the significance ($p<0.05$) of the most critical measurement for the model, the Spearman correlation [261] was computed between the values of each measurement across assessments with the corresponding predicted score. In parallel, the Mann–Whitney U test [184] examined the difference in the average measurement values between controls and the anhedonia (or negative valence) group. A high correlation between measurement and predictions and measurement and ground truth, would highlight whether the model was influenced by a meaningful feature.

7.3.6 Network Architecture and Optimization

The selected model architecture can be seen in Fig. 7.3. After a Fully Connected Layer (FC) [104], [204], the model was split into two branches. For the longitudinal prediction of the anhedonia or negative valence score, a Gated Recurrent Unit (GRU) [62] Layer was employed. A GRU was suitable for this task due to its capability of taking into account the information of all assessments provided for each subject and was not limited to processing one report at a time [126]. Furthermore, the low amount of trainable parameters in a GRU contributed towards combating overfitting.

Tab. 7.3. P-value of each category in predicting anhedonia and negative valence and the average difference and standard deviation in BACC when the category was permuted compared to the unpermuted data (i.e., BACC of 0.7513 for anhedonia and 0.7957 for negative valence). Categories are ranked with respect to anhedonia according to their p-values and difference if they had the same p-value. Bold marks p-values of significant importance (p<0.05).

	Anhedonia		**Negative Valence**	
Category	**p-value**	**Difference**	**p-value**	**Difference**
Personality	<**0.002**	-0.0753 ± 0.045	<**0.002**	-0.1517 ± 0.062
Life	**0.008**	-0.0233 ± 0.040	<**0.002**	-0.0427 ± 0.085
BRIEF	**0.008**	-0.0173 ± 0.061	<**0.002**	-0.0417 ± 0.106
Support	**0.010**	-0.0233 ± 0.063	0.432	-0.0007 ± 0.079
Sleep	**0.042**	-0.0133 ± 0.054	**0.024**	-0.0157 ± 0.093
Neuropsych	0.328	-0.0033 ± 0.055	0.562	+0.0013 ± 0.088
Substance Use	0.332	-0.0023 ± 0.048	**0.048**	-0.0097 ± 0.098
Demographics	0.672	+0.0017 ± 0.059	0.764	+0.0033 ± 0.096

Since the age of subjects varied across time-steps, the auxiliary task of age regression was utilized to encourage the model to align the reports across subjects implicitly. For the age-prediction branch, two FC layers were used. The final output of the model was the predicted score for anhedonia or negative valence per assessment, along with the regressed subject age.

The loss function used to train the status for anhedonia, and negative valence was a binary cross entropy loss $\mathcal{L}_{\text{BCE}}$ [204]. Since the dataset is characterized by class imbalance loss weighting was used to alleviate this problem. Specifically, the ratio of control to positive subjects, $R = \frac{N_{\text{control}}}{N_{\text{positive}}}$ was calculated, so that the loss would act as if the dataset contained $R \times N_{\text{positive}}$ positive subjects [78].

For the regression of the subject ages at each assessment, the Mean Squared Error Loss was selected $\mathcal{L}_{\text{MSE}}$. The model was trained with a composite loss function $\mathcal{L} = \alpha\mathcal{L}_{\text{BCE}} + (1-\alpha)\mathcal{L}_{\text{MSE}}$, where α moderated the contribution of each loss towards the overall optimization. In the experiments, α was empirically set to 0.8. Additionally, L1 weight regularization was employed to limit overfitting [307]. The model was trained for 30 epochs with learning rate 0.0001 and the Adam Optimizer [154] and was implemented in PyTorch [218]. Every batch contained all the assessments of one subject.

7.3.7 Results

The model's prediction accuracy was significant (p<0.001) for both constructs, i.e., the BACC was 75.13% for anhedonia and 79.57% for negative valence.

Tab. 7.4. The 3 most important measurements of categories crucial to predicting both constructs computed using guided backpropagation.

	Anhedonia	Negative Valence
Personality	Extraversion Emotional Stability Acceptance	Emotional Stability Extraversion Acceptance
Life	Negative Events Sexual Abuse Positive Events	Positive Events Sexual Abuse Chronic Negative Scale
BRIEF	Cognitive Shift T-score Behavioral Shift T-score Inhibit T-score	Cognitive Shift T-score Behavioral Shift T-score Inhibit T-score
Sleep	Trouble Sleeping Weekday Wake-up Weekday Sleep	Circadian Preference Weekend Sleep Trouble Sleeping

Of significant importance for predicting anhedonia, as can be seen in Table 7.3 were the categories personality ($p<0.002$), life ($p=0.008$), BRIEF ($p=0.008$), support ($p=0.01$), and sleep ($p=0.042$). For negative valence, categories of importance, as can be seen in Table 7.3 were personality ($p<0.002$), life ($p<0.002$), BRIEF ($p<0.002$), sleep ($p=0.024$), and substance use ($p=0.048$).

For the four categories that were of significant importance for predicting both constructs (i.e., personality, life, BRIEF, and sleep), the three most important measurements indicated by the guided backpropagation are listed in Table 7.4. Important for predicting either construct were the personality trades of extroversion, emotional stability, and acceptance. Aspects of life predicting both constructs were sexual abuse and positive events. BRIEF measurements important for both prediction tasks were the cognitive shift, inhibit, and behavioral shift T-scores. Finally, sleep disorder was important for predicting both anhedonia and negative valence.

For both anhedonia, as can be seen in Fig. 7.4 and negative valence in Fig. 7.5, the most critical measurements of each category significantly correlated ($p<0.05$) with the prediction score of the deep learning model (top) and reported ground truth status for each construct (bottom).

A higher prediction score was associated with lower extraversion for anhedonia, higher negative events, higher cognitive shift T-score, and increased sleep disorder. A higher negative valence score was associated with lower emotional stability, less positive controllable events, higher metacognition, and decreased circadian preference. Those findings for both constructs were confirmed when comparing controls with individuals reporting anhedonia or negative valence at least once.

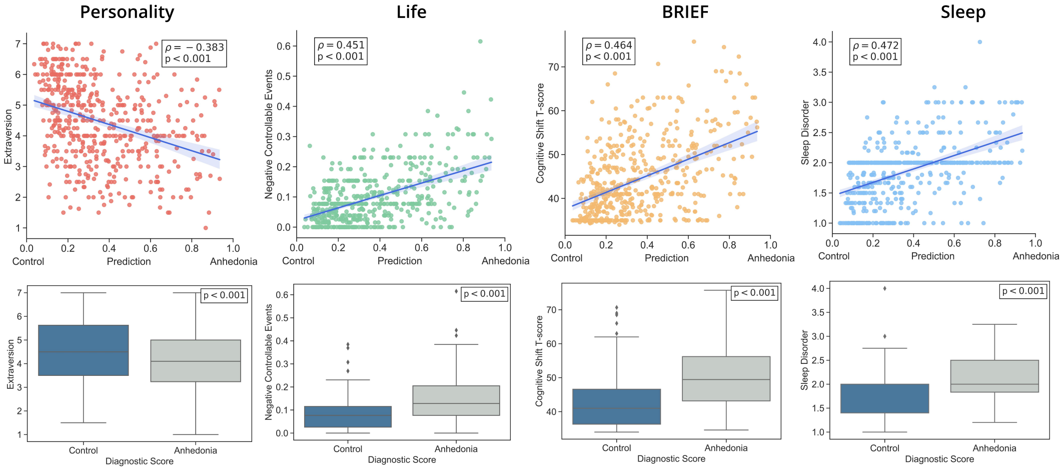


Fig. 7.4. Top: Correlation between the predicted score and the most important measurement value in each significant category for **anhedonia**. ρ denotes the Spearman's rank correlation coefficient. Bottom: Distribution of the most important measurements for controls with individuals reporting anhedonia at least in one assessment (medians ± Interquartile range and outliers). The predicted scores and measurements have been averaged across assessments.

7.3.8 Discussion

The longitudinal machine learning model accurately predicted the status of anhedonia (75.13% BACC) and negative valence (79.57% BACC) by also predicting the age of the subject at each assessment. Doing so, implicitly modeled differences in the characteristics of each construct across the adolescent age span.

The stability of the analysis was underlined by the consistent results of the permutation test for the two constructs, as can be seen in Table 7.3: 4 out of 8 categories were significantly important for both prediction tasks (i.e., personality, life, BRIEF, and sleep - in that order), while Neuropsych and Demographics did not have significant importance for either task. The only exception was Support in predicting anhedonia, which was, besides Substance use, the only category of significant importance for only one prediction task. Finally, the stability of our analysis was underlined by the significant correlations of the most essential measurements of each of the four significant categories with the prediction scores and the self-reported constructs as can be seen in Fig. 7.4 and Fig. 7.5.

Personality factors, particularly lower extraversion, lower emotional stability, and lower acceptance, were strong predictors of anhedonia. These findings are in line with previous literature showing a link between depression and extraversion [157], especially with the low positive emotionality component of extraversion [290].

Our model identified negative life events in general (controllable negative events for anhedonia, chronic events for negative valence), childhood sexual abuse, and low positive events as predictors of both anhedonia and negative valence. Research has consistently documented increased susceptibility to depression in youth with more adverse life events [264]. It has specifically been well documented [148] that sexual abuse has a strong association with depression. Adolescents might differ in their reactions to life events based on their control

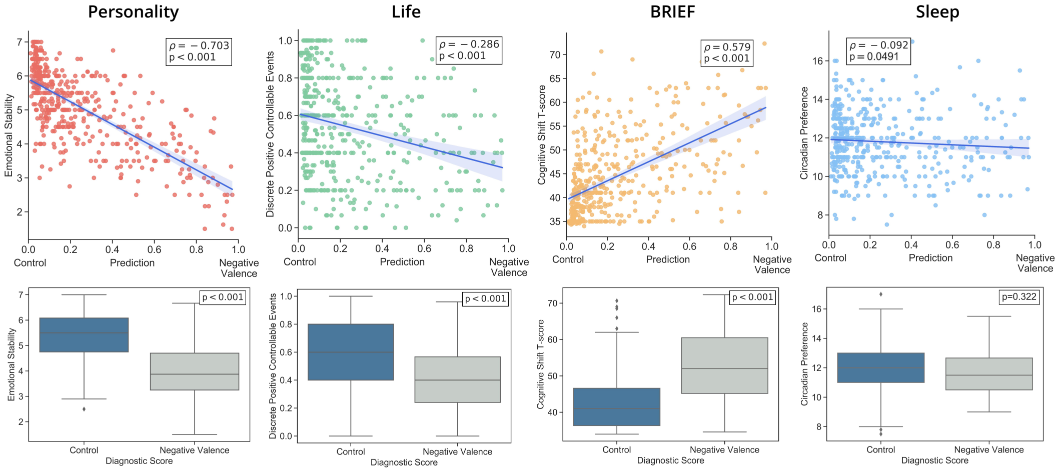


Fig. 7.5. Top: Correlation between the predicted score and the most important measurement value in each significant category for **negative valence**. ρ denotes the Spearman's rank correlation coefficient. Bottom: Distribution of the most important measurements for controls with individuals reporting negative valence at least in one assessment (medians ± Interquartile range and outliers). The predicted scores and measurements have been averaged across assessments.

in the situation. Life events that fall beyond individual control are labeled as uncontrollable events, while events influenced by the individuals are referred to as and controllable life events [190]. Our results highlighting the effect of controllable negative life events confirm the results of others [137], who argue that controllable events are more likely to increase the likelihood of psychiatric morbidity and exacerbate the symptom levels over time.

Moreover, our results show a consistent association between depressive symptoms and executive dysfunction. Namely, higher inhibitory control and lower flexibility (cognitive and behavioral shift) are associated with both anhedonia and negative valence. Impaired executive control over negative information may lead to increased negative cognitions and prolonged negative affect, increasing the risk for depression. Depressed subjects recognize happy facial expressions more slowly [269] than controls. Subsequent studies have similarly noted that adolescents with a recent first episode of major depression show attention shift towards sad stimuli and more impulsive behavior in decision making [165]. These results suggest that specific patterns of neuropsychological functions may be impacted selectively in the first episode of major depression [165].

Current results indicate the importance of sleep behavior as a predictor of both anhedonia and negative valence. Specifically, poor sleep quality, frequent awakenings, and shortened sleep duration predicted anhedonia and circadian preference towards morningness along with shorter sleep duration. Disturbed sleep predicted negative valence. During pubertal development, adolescents tend to move towards later chronotypes [235], and their sleep time is highly variable [94] which puts them uniquely at risk of sleep problems. Adolescents with shorter sleep duration assessed by daily self-report measures report greater depression, anxiety, fatigue, and lower subjective well-being controlling for average sleep duration [94].

The results confirm the previously documented circadian preference towards eveningness being related to poorer mental health and higher prevalence of clinical depression [156].

Given that sleep is a modifiable factor [30], improving sleep quality could be an important intervention strategy to avoid the development of mood problems during adolescence. Studies suggest that sleep has an active role in brain maturation [275]. Our data support sleep as a protective factor against the emergence of depressive symptoms, which could ultimately translate into reduced risk for depressive disorder.

Identifying the developmental characteristics is out of the scope of the current analysis; however, previous results show that anhedonia stabilizes over adolescence [25]. The lack of social support appeared to be important in anhedonia suggesting supportive, and age congruent interpersonal relationships contribute to the resilience of youth to anhedonia specifically.

Substance use was a risk factor for negative valence. There is a bidirectional relationship between heavy alcohol use and depression, with shared risk factors; alcohol may be used to relieve negative feelings, but alcohol problems can also predispose people to depression [189]. A meta-analysis of several studies showed that more frequent engagement in alcohol use and binge drinking are associated with higher levels of depression in adolescents [45].

Sex was not a significant predictor of either anhedonia or negative valence. Sex differences in depression are well documented, being more common in women, beginning during adolescence [112], and reflected in the higher proportion of girls in the anhedonia and negative valence groups. However, we intentionally made the models sensitive to the developmental transitions by partially predicting age, which likely reduced sensitivity to any sex effect.

Our experiments showed that training and analyzing a model using solely tabular data successfully predicted anhedonia and negative valence. Moreover, even though model interpretation methods, such as guided backpropagation, have been proposed for imaging-based models, we showed that they could be applied to tabular data as well.

Future work includes incorporating network-based brain activation patterns based on functional or resting-state MRI data to determine whether our current results based on self-reported measures are reflected in the subsequent neural substrates. Studies [197] have highlighted that high-risk adolescents are characterized by altered cortical thickness in regions of the brain involved in cognitive control, emotional regulation, and default mode networks, and suggest that alternative modeling, focusing on the underlying neural representation may provide additional insights about the development of depression and characterization of anhedonia and negative valence in adolescents [9, 90].

Overall our permutation testing identified categories highly associated with adolescent depression. The proposed pipeline achieved interpretable and meaningful predictions and gave insights into which factors and measurements contribute the most to two constructs of adolescent major depressive disorder.

Part V

Conclusions

Conclusions

8

8.1 Summary and Findings

In this dissertation, we discussed the concept of robustness in deep models for medical imaging applications. In this Chapter, we will provide an overall summary for each work and discuss future outlooks.

8.1.1 Robustness Improvement

Chapter 3 provided an overview of adversarial attack crafting methods and their applications to semantic segmentation, detection, and speech recognition. Furthermore, a taxonomy of adversarial defenses was described, and the most prominent defense methods were discussed. Finally, we showed how adversarial examples have been used so far in the context of medical imaging.

Chapter 4 described a novel data augmentation method using affine geometric transformations and quantified the robustness of machine learning models. Extensive experiments on medical imaging diagnostic tasks, namely fine-grained skin lesion classification and mammogram tumor classification, highlighted the advantages of ManiFool Augmentation. Models trained with the proposed augmentation outperformed other data augmentation approaches on the clean test set. Moreover, the robustness of the models trained with ManiFool Augmentation was drastically increased both for random affine and projective transformations. Experiments across datasets in an unseen test scenario also underlined the increased capabilities of the models trained with ManiFool augmentation.

Additionally, a quantitative metric for the robustness of machine learning models was computed based on the geodesic distance of the clean samples to the model decision boundaries. Our experiments showed how different state-of-the-art architectures could achieve various levels of robustness for different datasets. The proposed augmentation method and metric are general and could be applied to various imaging diagnostic tasks for different modalities.

8.1.2 Enhanced Training Dynamics

In Chapter 5, we introduced 3DQ, a ternary quantization technique that can achieve 16x model compression. For the first time, our method was applied to 3D F-CNNs that performed volumetric whole-brain and hippocampus segmentation of MRI scans. Our experiments validated that the models quantized with 3DQ performed equally well or better than the baselines, including the full precision networks, for two medical imaging datasets.

Comparing our model to other compression techniques, such as knowledge distillation, showcased the increased performance of 3DQ. Finally, for 3DU-Net, our approach outperformed the full precision models for both tasks, highlighting that 3DQ can limit overfitting and enhance training dynamics when training large models with limited data. Finally, due to 16x network compression, 3DQ constitutes a solid approach for space-critical applications, like patient-specific models or model weight transfer for Federated Learning.

8.1.3 Model Benchmarking with Adversarial Examples

In Chapter 6, we discussed methods that evaluate the model robustness, namely model testing and verification. We described the most commonly used and practical model benchmarking datasets and the limitations of current model verification techniques. Afterwards, we described our contribution, benchmarking models for medical image classification and segmentation using task-specific adversarial attacks. We evaluated 3 commonly used network architectures for each task using 3 different attack mechanisms in a black-box scenario.

Our experiments found that models with similar performance on clean data have notable differences in their relative exploration of the underlying data manifold, resulting in varying robustness capabilities. Specifically, we illustrated that for segmentation tasks, dense blocks and skip connections contributed to enhanced generalizability and robustness, while model depth seemed to increase the resilience of classification models to adversarial examples.

8.1.4 Robustness of Non-Imaging Data

In Chapter 7, we discussed the concept of robustness beyond imaging data. Specifically, we introduced the challenges of crafting adversarial attacks for tabular data and different ways that imaging and non-imaging data can be fused before being processed by a DNN. Afterwards, we discussed how major depressive disorder can be increased during adolescence and that there is a need to identify the risk and protective factors of depression in youth.

Moreover, we designed a novel type of mental health prediction in adolescents using a multi-task recurrent neural network. Based on individual histories, we identified risk factors for symptoms of depression, specifically anhedonia and negative valence, from a large pool of cognitive, emotional and personality factors, considered in the context of developmental changes occurring across this age span.

In order to evaluate our model, we performed permutation testing and identified the most significant variable categories that contributed to the model's decision. Afterwards, we used model interpretation techniques to rank the importance of individual input measurements. Finally, we correlated the top-ranked measurements with the predicted scores and the ground truth to evaluate their significance.

8.2 Future Outlook

This dissertation focused on methods to improve and evaluate the robustness in the context of supervised learning. However, curating and annotating large-scale datasets is time-consuming, costly, and in some cases, infeasible. Therefore, robustness investigation for unsupervised learning is an interesting and valuable direction.

Furthermore, identifying the types of robustness most relevant for every medical application is key. Some of the most common challenges faced by computer-aided systems are dataset shift and outliers or anomalies. Thus, ensuring models are resilient against these circumstances will highly benefit DNNs for medical applications. Moreover, developing accurate outlier or adversarial example detection systems incorporating model uncertainty and interpretation will lower the risk of test-time threats.

Furthermore, incorporating additional information in the form of anatomical context or tabular data is a promising direction to enhance the performance and robustness of DNNs for medical applications. Combining tabular and imaging data is an open field of study, where novel model architectures and aggregation schemes could be very beneficial.

Overall, thoroughly evaluating the performance of a DNN beyond an unseen test set is extremely critical. As we showed, models with similar performances on clean data can vary significantly in terms of robustness. Therefore, designing an evaluation pipeline that includes model interpretation, statistical analysis, and a diverse test-set of potential inputs that could cause model failure is highly recommended. Similarly, developing model verification techniques that will overcome the discussed limitations and provide guarantees for a model's performance will be a large step forward in deep learning security.

We hope this dissertation will inspire future research towards building more robust machine learning systems for healthcare and crafting thorough evaluation pipelines that analyze and interpret the model's decisions.

Part VI

Appendix

Authored and Co-authored Publications

A

Authored

1. **M. Paschali**, W. Simson, A. Guha Roy, R. Göbl, C. Wachinger, N. Navab. *"Manifold Exploring Data Augmentation with Geometric Transformations for Increased Performance and Robustness."* International Conference on Information Processing in Medical Imaging (IPMI), Hong Kong, 2019

2. **M. Paschali**, S. Conjeti, F. Navarro, N. Navab. *"Generalizability vs. Robustness: Adversarial Examples for Medical Imaging."* International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Granada, 2018

3. **M. Paschali***, S. Gasperini*, A. Guha Roy, M.Y.-S. Fang, N. Navab. *"3DQ: Compact Quantized Neural Networks for Volumetric Whole Brain Segmentation."* International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Shenzhen, 2019 (Equal Contribution)

4. **M. Paschali**, O. Kiss, Q. Zhao, E. Adeli, S. Podhajsky, I. Gotlib, K.M. Pohl, F. Baker *"Predicting Symptoms of Depression in Adolescents based on Longitudinal Self-Reports and Behavioral Assessments."*, 2021 (Under Review)

5. **M. Paschali***, M.F. Naeem*, W. Simson, K. Steiger, M. Mollenhauer, N. Navab. *"Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks."* arXiv preprint arXiv:1904.03127, 2019 (Equal Contribution)

Co-authored

1. T. Czempiel, **M. Paschali**, D. Ostler, S.T. Kim, B. Busam, N. Navab. *"OperA: Attention-Regularized Transformers for Surgical Phase Recognition."* International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

2. S.T. Kim, L. Goli, **M. Paschali**, A. Khakzar, M. Keicher, T. Czempiel, E. Burian, R. Braren, N. Navab, T. Wendler. *"Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs."* International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

3. M. Tirindelli*, C. Eilers*, W. Simson, **M. Paschali**, M.F. Azampour, N. Navab. *"Rethinking Ultrasound Augmentation: A Physics-Inspired Approach."* International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021 (Equal Contribution)

4. C. Berger, **M. Paschali**, B. Glocker, K. Kamnitsas. Confidence-based *"Out-of-Distribution Detection: A Comparative Study and Analysis."* International Conference on Medical Image Computing and Computer Assisted Intervention Workshop - Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE), 2021

5. M. Keicher*, H. Burwinkel*, D. Bani-Harouni*, **M. Paschali**, T. Czempiel, E. Burian, M.R. Makowski, R. Braren, N. Navab, T. Wendler. *"U-GAT: Multimodal Graph Attention Network for COVID-19 Outcome Prediction."* arXiv preprint arXiv:2108.00860, 2021 (Equal Contribution)

6. W. Simson, **M. Paschali**, V. Sideri-Lampretsa, N. Navab, J.J. Dahl. *"Investigating Pulse-Echo Sound Speed Estimation in Breast Ultrasound with Deep Learning."*, 2021 (Under Submission)

7. M.X. Foo, S.T. Kim, **M. Paschali**, L. Goli, E. Burian, M. Makowski, R. Braren, N. Navab, T. Wendler. *"Interactive Segmentation for COVID-19 Infection Quantification on Longitudinal CT scans."*, arXiv preprint arXiv:2110.00948, 2021

8. T. Czempiel, **M. Paschali**, M. Keicher, W. Simson, H. Feussner, S.T. Kim, N. Navab. *"TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks."* International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Lima, 2020

9. H. Hase*, M.F. Azampour*, M. Tirindelli, **M. Paschali**, W. Simson, E. Fatemizadeh, N. Navab. *"Ultrasound-Guided Robotic Navigation with Deep Reinforcement Learning."* IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, 2020 (Equal Contribution)

10. S. Gasperini, **M. Paschali**, C. Hopke, D. Wittmann, N. Navab. *"Signal Clustering with Class-independent Segmentation."* International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, 2020

11. W. Simson, R. Göbl, **M. Paschali**, M. Krönke, K. Scheidhauer, W. Weber, N. Navab. *"End-to-End Learning-Based Ultrasound Reconstruction."* arXiv preprint arXiv/1904.04696, 2019

12. P. Notaro, **M. Paschali**, C. Hopke, D. Wittmann, N. Navab. *"Radar Emitter Classification with Attribute-specific Recurrent Neural Networks."* arXiv preprint arXiv/1911.07683, 2019

13. W. Simson, **M. Paschali**, N. Navab, G. Zahnd. *"Deep learning beamforming for sub-sampled ultrasound data."* IEEE International Ultrasonics Symposium (IUS), Kobe, 2018

B Abstracts of Publications not Discussed in this Thesis

OperA: Attention-Regularized Transformers for Surgical Phase Recognition

T. Czempiel, **M. Paschali**, D. Ostler, S.T. Kim, B. Busam, N. Navab. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

Abstract used with permission from Springer Nature Customer Service Centre GmbH with license number: 5175920414050

In this paper we introduce OperA, a transformer-based model that accurately predicts surgical phases from long video sequences. A novel attention regularization loss encourages the model to focus on high-quality frames during training. Moreover, the attention weights are utilized to identify characteristic high attention frames for each surgical phase, which could further be used for surgery summarization. OperA is thoroughly evaluated on two datasets of laparoscopic cholecystectomy videos, outperforming various state-of-the-art temporal refinement approaches.

Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs

S.T. Kim, L. Goli, **M. Paschali**, A. Khakzar, M. Keicher, T. Czempiel, E. Burian, R. Braren, N. Navab, T. Wendler. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

Abstract used with permission from Springer Nature Customer Service Centre GmbH with license number: 5175920548475

Chest computed tomography (CT) has played an essential diagnostic role in assessing patients with COVID-19 by showing disease-specific image features such as ground-glass opacity and consolidation. Image segmentation methods have proven to help quantify the disease burden and even help predict the outcome. The availability of longitudinal CT series may also result in an efficient and effective method to reliably assess the progression of COVID-19, monitor the healing process and the response to different therapeutic strategies. In this paper, we propose a new framework to identify infection at a voxel level (identification of healthy lung, consolidation, and ground-glass opacity) and visualize the progression of COVID-19 using sequential low-dose non-contrast CT scans. In particular, we devise a longitudinal segmentation network that utilizes the reference scan information to improve the performance of disease identification. Experimental results on a clinical longitudinal dataset collected in our institution show the effectiveness of the proposed method compared to the static deep neural networks for disease quantification.

Rethinking Ultrasound Augmentation: A Physics-Inspired Approach

M. Tirindelli*, C. Eilers*, W. Simson, **M. Paschali**, M.F. Azampour, N. Navab. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021 (Equal Contribution)

Abstract used with permission from Springer Nature Customer Service Centre GmbH with license number: 5175920621524

Medical Ultrasound (US), despite its wide use, is characterized by artifacts and operator dependency. Those attributes hinder the gathering and utilization of US datasets for the training of Deep Neural Networks used for Computer-Assisted Intervention Systems. Data augmentation is commonly used to enhance model generalization and performance. However, common data augmentation techniques, such as affine transformations do not align with the physics of US and, when used carelessly can lead to unrealistic US images. To this end, we propose a set of physics-inspired transformations, including deformation, reverb and Signal-to-Noise Ratio, that we apply on US B-mode images for data augmentation. We evaluate our method on a new spine US dataset for the tasks of bone segmentation and classification.

Confidence-based Out-of-Distribution Detection: A Comparative Study and Analysis.

C. Berger, **M. Paschali**, B. Glocker, K. Kamnitsas. International Conference on Medical Image Computing and Computer Assisted Intervention Workshop - Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE), 2021

Abstract used with permission from Springer Nature Customer Service Centre GmbH with license number: 5175920757328

Image classification models deployed in the real world may receive inputs outside the intended data distribution. For critical applications such as clinical decision making, it is important that a model can detect such out-of-distribution (OOD) inputs and express its uncertainty. In this work, we assess the capability of various state-of-the-art approaches for confidence-based OOD detection through a comparative study and in-depth analysis. First, we leverage a computer vision benchmark to reproduce and compare multiple OOD detection methods. We then evaluate their capabilities on the challenging task of disease classification using chest X-rays. Our study shows that high performance in a computer vision task does not directly translate to accuracy in a medical imaging task. We analyse factors that affect performance of the methods between the two tasks. Our results provide useful insights for developing the next generation of OOD detection methods.

U-GAT: Multimodal Graph Attention Network for COVID-19 Outcome Prediction

M. Keicher*, H. Burwinkel*, D. Bani-Harouni*, **M. Paschali**, T. Czempiel, E. Burian, M.R. Makowski, R. Braren, N. Navab, T. Wendler. arXiv preprint arXiv:2108.00860, 2021 (Equal Contribution)

During the first wave of COVID-19, hospitals were overwhelmed with the high number of admitted patients. An accurate prediction of the most likely individual disease progression can improve the planning of limited resources and finding the optimal treatment for patients. However, when dealing with a newly emerging disease such as COVID-19, the impact of patient- and disease-specific factors (e.g. body weight or known co-morbidities) on the immediate course of disease is by and large unknown. In the case of COVID-19, the need for intensive care unit (ICU) admission of pneumonia patients is often determined only by acute indicators such as vital signs (e.g. breathing rate, blood oxygen levels), whereas statistical analysis and decision support systems that integrate all of the available data could enable an earlier prognosis. To this end, we propose a holistic graph-based approach combining both imaging and non-imaging information. Specifically, we introduce a multimodal similarity metric to build a population graph for clustering patients and an image-based end-to-end Graph Attention Network to process this graph and predict the COVID-19 patient outcomes: admission to ICU, need for ventilation and mortality. Additionally, the network segments chest CT images as an auxiliary task and extracts image features and radiomics for feature fusion with the available metadata. Results on a dataset collected in Klinikum rechts der Isar in Munich, Germany show that our approach outperforms single modality and non-graph baselines. Moreover, our clustering and graph attention allow for increased understanding of the patient relationships within the population graph and provide insight into the network's decision-making process.

Investigating Pulse-Echo Sound Speed Estimation in Breast Ultrasound with Deep Learning

W. Simson, **M. Paschali**, V. Sideri-Lampretsa, N. Navab, J.J. Dahl. Under Submission, 2021

Ultrasound is an adjunct tool to mammography that can quickly and safely aid physicians with diagnosing breast abnormalities. In clinical ultrasound, a constant speed of sound is used to form the B-mode images for diagnosis. However, the various types of breast tissue, such as glandular, fat, and lesions, differ in speed of sound. These differences can degrade the image reconstruction process. Alternatively, speed of sound can be utilized as a powerful tool for identifying disease. To this end, we propose a deep-learning approach for sound speed estimation from IQ ultrasound signals. First, we develop a large-scale simulated ultrasound dataset that approximates realistic breast tissue which models breast gland, skin and lesions with varying echogenicity and speed of sound. We developed a fully convolutional network architecture that is trained with the simulated dataset to produce an estimated map of the speed of sound from the input of three complex-valued IQ ultrasound images formed from plane-wave transmissions at separate angles. Furthermore, thermal noise augmentation is used during model optimization to enhance generalizability to real ultrasound data. Our model is extensively evaluated on simulated, phantom and in-vivo breast ultrasound data, demonstrating its ability to accurately estimate sound speeds that are consistent with previously reported values in the literature. Our simulated dataset and model will become publicly available to provide a step towards accurate and generalizable sound speed estimation for pulse-echo ultrasound imaging.

Interactive Segmentation for COVID-19 Infection Quantification on Longitudinal CT scans

M.X. Foo, S.T. Kim, **M. Paschali**, L. Goli, E. Burian, M. Makowski, R. Braren, N. Navab, T. Wendler. arXiv preprint arXiv:2110.00948, 2021

Consistent segmentation of COVID-19 patient's CT scans across multiple time points is essential to assess disease progression and response to therapy accurately. Existing automatic and interactive segmentation models for medical images only use data from a single time point (static). However, valuable segmentation information from previous time points is often not used to aid the segmentation of a patient's follow-up scans. Also, fully automatic segmentation techniques frequently produce results that would need further editing for clinical use. In this work, we propose a new single network model for interactive segmentation that fully utilizes all available past information to refine the segmentation of follow-up scans. In the first segmentation round, our model takes 3D volumes of medical images from two-time points (target and reference) as concatenated slices with the additional reference time point segmentation as a guide to segment the target scan. In subsequent segmentation refinement rounds, user feedback in the form of scribbles that correct the segmentation and the target's previous segmentation results are additionally fed into the model. This ensures that the segmentation information from previous refinement rounds is retained. Experimental results on our in-house multiclass longitudinal COVID-19 dataset show that the proposed model outperforms its static version and can assist in localizing COVID-19 infections in patient's follow-up scans.

TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks

T. Czempiel, **M. Paschali**, M. Keicher, W. Simson, H. Feussner, S.T. Kim, N. Navab. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Lima, 2020

Automatic surgical phase recognition is a challenging and crucial task with the potential to improve patient safety and become an integral part of intra-operative decision-support systems. In this paper, we propose, for the first time in workflow analysis, a Multi-Stage Temporal Convolutional Network (MS-TCN) that performs hierarchical prediction refinement for surgical phase recognition. Causal, dilated convolutions allow for a large receptive field and online inference with smooth predictions even during ambiguous transitions. Our method is thoroughly evaluated on two datasets of laparoscopic cholecystectomy videos with and without the use of additional surgical tool information. Outperforming various state-of-the-art LSTM approaches, we verify the suitability of the proposed causal MS-TCN for surgical phase recognition.

Ultrasound-Guided Robotic Navigation with Deep Reinforcement Learning

H. Hase*, M.F. Azampour*, M. Tirindelli, **M. Paschali**, W. Simson, E. Fatemizadeh, N. Navab. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, 2020 (Equal Contribution)

In this paper we introduce the first reinforcement learning (RL) based robotic navigation method which utilizes ultrasound (US) images as an input. Our approach combines state-of-the-art RL techniques, specifically deep Q-networks (DQN) with memory buffers and a binary classifier for deciding when to terminate the task. Our method is trained and evaluated on an in-house collected data-set of 34 volunteers and when compared to pure RL and supervised learning (SL) techniques, it performs substantially better, which highlights the suitability of RL navigation for US-guided procedures. When testing our proposed model, we obtained a 82.91% chance of navigating correctly to the sacrum from 165 different starting positions on 5 different unseen simulated environments.

Signal Clustering With Class-Independent Segmentation

S. Gasperini, **M. Paschali**, C. Hopke, D. Wittmann, N. Navab. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, 2020

Radar signals have been dramatically increasing in complexity, limiting the source separation ability of traditional approaches. In this paper we propose a Deep Learning-based clustering method, which encodes concurrent signals into images, and, for the first time, tackles clustering with image segmentation. Novel loss functions are introduced to optimize a Neural Network to separate the input pulses into pure and non-fragmented clusters. Outperforming a variety

of baselines, the proposed approach is capable of clustering inputs directly with a Neural Network, in an end-to-end fashion.

Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks

M. Paschali*, M.F. Naeem*, W. Simson, K. Steiger, M. Mollenhauer, N. Navab. arXiv preprint arXiv:1904.03127, 2019 (Equal Contribution)

In this paper, we propose a novel interpretation method tailored to histological Whole Slide Image (WSI) processing. A Deep Neural Network (DNN), inspired by Bag-of-Features models is equipped with a Multiple Instance Learning (MIL) branch and trained with weak supervision for WSI classification. MIL avoids label ambiguity and enhances our model's expressive power without guiding its attention. We utilize a fine-grained logit heatmap of the models activations to interpret its decision-making process. The proposed method is quantitatively and qualitatively evaluated on two challenging histology datasets, outperforming a variety of baselines. In addition, two expert pathologists were consulted regarding the interpretability provided by our method and acknowledged its potential for integration into several clinical applications.

Deep learning beamforming for sub-sampled ultrasound data

W. Simson, **M. Paschali**, N. Navab, G. Zahnd. IEEE International Ultrasonics Symposium (IUS), Kobe, 2018

In medical imaging tasks, such as cardiac imaging, ultrasound acquisition time is crucial, however traditional high-quality beamforming techniques are computationally expensive and their performance is hindered by sub-sampled data. To this end, we propose DeepFormer, a method to reconstruct high quality ultrasound images in real-time on sub-sampled raw data by performing an end-to-end deep learning-based reconstruction. Results on an in vivo dataset of 19 participants show that DeepFormer offers promising advantages over traditional processing of sub-sampled raw-ultrasound data and produces reconstructions that are both qualitatively and visually equivalent to fully-sampled DeepFormed images.

End-to-end learning-based ultrasound reconstruction

W. Simson, R. Göbl, **M. Paschali**, M. Krönke, K. Scheidhauer, W. Weber, N. Navab. arXiv preprint arXiv/1904.04696, 2019

Ultrasound imaging is caught between the quest for the highest image quality, and the necessity for clinical usability. Our contribution is two-fold: First, we propose a novel fully convolutional neural network for ultrasound reconstruction. Second, a custom loss function tailored to the modality is employed for end-to-end training of the network. We demonstrate that training a network to map time-delayed raw data to a minimum variance ground truth offers performance increases in a clinical environment. In doing so, a path is explored towards improved clinically viable ultrasound reconstruction. The proposed method displays both promising image reconstruction quality and acquisition frequency when integrated for live ultrasound scanning. A clinical evaluation is conducted to verify the diagnostic usefulness of the proposed method in a clinical setting.

Radar Emitter Classification with Attribute-specific Recurrent Neural Networks

P. Notaro, **M. Paschali**, C. Hopke, D. Wittmann, N. Navab. arXiv preprint arXiv/1911.07683, 2019

Radar pulse streams exhibit increasingly complex temporal patterns and can no longer rely on a purely value-based analysis of the pulse attributes for the purpose of emitter classification. In this paper, we employ Recurrent Neural Networks (RNNs) to efficiently model and exploit the temporal dependencies present inside pulse streams. With the purpose of enhancing the network prediction capability, we introduce two novel techniques: a per-sequence normalization, able to mine the useful temporal patterns; and attribute-specific RNN processing, capable of processing the extracted information effectively. The new techniques are evaluated with an ablation study and the proposed solution is compared to previous Deep Learning (DL) approaches. Finally, a comparative study on the robustness of the same approaches is conducted and its results are presented.

Bibliography

[1] *A Funding Crisis for Public Health and Safety*. `https://www.tfah.org/report-details/a-funding-crisis-for-public-health-and-safety-state-by-state-and-federal-public-health-funding-facts-and-recommendations/` (cit. on p. 4).

[2] M. Abadi, A. Agarwal, P. Barham, et al. “Tensorflow: Large-scale machine learning on heterogeneous distributed systems”. In: *arXiv preprint arXiv:1603.04467* (2016) (cit. on p. 67).

[3] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009 (cit. on p. 34).

[4] T. Achenback and L. Rescorla. “Manual for the ASEBA school-age forms & profiles”. In: *Burlington, VT: University of Vermont Research Centre for Children, Youth and Families* (2001) (cit. on p. 80).

[5] *Adverse Childhood Experiences International Questionnaire (ACE-IQ), World Health Organisation*. `https://www.who.int/violence_injury_prevention/violence/activities/adverse_childhood_experiences/en/`. Accessed: 2021-05-08 (cit. on p. 82).

[6] M. Aghili, S. Tabarestani, M. Adjouadi, and E. Adeli. “Predictive modeling of longitudinal data for Alzheimer’s Disease Diagnosis Using RNNs”. In: *International Workshop on PRedictive Intelligence In MEdicine*. Springer. 2018, pp. 112–119 (cit. on pp. 76, 77).

[7] M. Alzantot, B. Balaji, and M. Srivastava. “Did you hear that? adversarial examples against automatic speech recognition”. In: *arXiv preprint arXiv:1801.00554* (2018) (cit. on p. 18).

[8] S. An, C. Xiao, W. F. Stewart, and J. Sun. “Longitudinal adversarial attack on electronic health records data”. In: *The World Wide Web Conference*. 2019, pp. 2558–2564 (cit. on p. 78).

[9] S. L. Andersen and M. H. Teicher. “Stress, sensitive periods and maturational events in adolescent depression”. In: *Trends in neurosciences* 31.4 (2008), pp. 183–191 (cit. on p. 92).

[10] A. Antoniou, A. Storkey, and H. Edwards. “Data augmentation generative adversarial networks”. In: *arXiv preprint arXiv:1711.04340* (2017) (cit. on p. 32).

[11] S. Arlot, A. Celisse, et al. “A survey of cross-validation procedures for model selection”. In: *Statistics surveys* 4 (2010), pp. 40–79 (cit. on p. 85).

[12] A. Athalye and N. Carlini. “On the robustness of the cvpr 2018 white-box adversarial example defenses”. In: *arXiv preprint arXiv:1804.03286* (2018) (cit. on p. 23).

[13] A. Athalye, N. Carlini, and D. Wagner. “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 274–283 (cit. on pp. 19, 23).

[14] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. “Synthesizing robust adversarial examples”. In: *International conference on machine learning*. PMLR. 2018, pp. 284–293 (cit. on pp. 21, 22).

[15] S. Avenevoli, E. Knight, R. C. Kessler, and K. R. Merikangas. “Epidemiology of depression in children and adolescents.” In: (2008) (cit. on p. 78).

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495 (cit. on p. 66).

[17] T. Bai, J. Zhao, J. Zhu, et al. “Ai-gan: Attack-inspired generation of adversarial examples”. In: *arXiv preprint arXiv:2002.02196* (2020) (cit. on p. 17).

[18] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees. “A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions”. In: *Color Medical Image Analysis*. Springer, 2013, pp. 63–86 (cit. on pp. 37, 41).

[19] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees. “A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions”. In: *Color Medical Image Analysis*. Springer, 2013, pp. 63–86 (cit. on p. 66).

[20] V. Ballet, X. Renard, J. Aigrain, T. Laugel, P. Frossard, and M. Detyniecki. “Imperceptible adversarial attacks on tabular data”. In: *arXiv preprint arXiv:1911.03274* (2019) (cit. on p. 78).

[21] M. Balunovic and M. Vechev. “Adversarial training and provable defenses: Bridging the gap”. In: *International Conference on Learning Representations*. 2019 (cit. on p. 24).

[22] P. Bandi, O. Geessink, Q. Manson, et al. “From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge”. In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 550–560 (cit. on pp. 57, 58).

[23] T. P. Beauchaine, D. N. Klein, E. Knapton, and A. Zisner. “Anhedonia in Depression: Mechanisms, Assessment, and Therapeutics”. In: *Neurobiology of Depression*. Elsevier, 2019, pp. 31–41 (cit. on p. 80).

[24] V. Behzadan and A. Munir. “Vulnerability of deep reinforcement learning to policy induction attacks”. In: *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer. 2017, pp. 262–275 (cit. on p. 19).

[25] E. C. Bennik, E. Nederhof, J. Ormel, and A. J. Oldehinkel. “Anhedonia and depressed mood in adolescence: course, stability, and reciprocal relation in the TRAILS study”. In: *European child & adolescent psychiatry* 23.7 (2014), pp. 579–586 (cit. on pp. 79, 80, 92).

[26] D. P. Bernstein, L. Fink, L. Handelsman, et al. “Initial reliability and validity of a new retrospective measure of child abuse and neglect.” In: *The American journal of psychiatry* (1994) (cit. on p. 82).

[27] N. Bhagwat, J. D. Viviano, A. N. Voineskos, M. M. Chakravarty, A. D. N. Initiative, et al. “Modeling and prediction of clinical symptom trajectories in Alzheimer’s disease using longitudinal data”. In: *PLoS computational biology* 14.9 (2018), e1006376 (cit. on p. 76).

[28] B. Biggio, B. Nelson, and P. Laskov. “Poisoning attacks against support vector machines”. In: *arXiv preprint arXiv:1206.6389* (2012) (cit. on p. 14).

[29] B. Biggio and F. Roli. “Wild patterns: Ten years after the rise of adversarial machine learning”. In: *Pattern Recognition* 84 (2018), pp. 317–331 (cit. on p. 3).

[30] M. J. Blake, J. A. Trinder, and N. B. Allen. “Mechanisms underlying the association between insomnia, anxiety, and depression in adolescence: implications for behavioral sleep interventions”. In: *Clinical psychology review* 63 (2018), pp. 25–40 (cit. on p. 92).

[31] A. Bojchevski and S. Günnemann. “Adversarial attacks on node embeddings via graph poisoning”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 695–704 (cit. on p. 19).

[32] W. W. Boonn and C. P. Langlotz. “Radiologist use of and perceived need for patient data access”. In: *Journal of digital imaging* 22.4 (2009), pp. 357–362 (cit. on p. 76).

[33] C. Bowles, L. Chen, R. Guerrero, et al. “Gan augmentation: Augmenting training data using generative adversarial networks”. In: *arXiv preprint arXiv:1810.10863* (2018) (cit. on pp. 33, 39).

[34] C. Bowles, R. Gunn, A. Hammers, and D. Rueckert. “GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation”. In: *arXiv preprint arXiv:1811.10669* (2018) (cit. on p. 33).

[35] *Breach portal: Cases Currently Under Investigation*. `https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf`. Accessed: 2021-05-06 (cit. on p. 5).

[36] *Break neural networks in your browser*. `https://kennysong.github.io/adversarial.js/`. Accessed: 2021-05-03 (cit. on p. 16).

[37] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. “The balanced accuracy and its posterior distribution”. In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 3121–3124 (cit. on p. 85).

[38] S. Brown, T. Brumback, K. Tomlinson, et al. “The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA): characterizing risk and resilience for alcohol use in adolescents”. In: *J Stud Alcohol Drugs* 76 (2015), pp. 895–908 (cit. on pp. 79, 82).

[39] S. A. Brown, T. Brumback, K. Tomlinson, et al. “The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA): a multisite study of adolescent development and substance use”. In: *Journal of studies on alcohol and drugs* 76.6 (2015), pp. 895–908 (cit. on p. 83).

[40] S. A. Brown, M. G. Myers, L. Lippke, S. F. Tapert, D. G. Stewart, and P. W. Vik. “Psychometric evaluation of the Customary Drinking and Drug Use Record (CDDR): a measure of adolescent alcohol and drug involvement.” In: *Journal of studies on alcohol* 59.4 (1998), pp. 427–438 (cit. on p. 81).

[41] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. “Adversarial patch”. In: *arXiv preprint arXiv:1712.09665* (2017) (cit. on p. 16).

[42] K. K. Bucholz, R. Cadoret, C. R. Cloninger, et al. “A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA.” In: *Journal of studies on alcohol* 55.2 (1994), pp. 149–158 (cit. on p. 82).

[43] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer. “The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research”. In: *Psychiatry research* 28.2 (1989), pp. 193–213 (cit. on p. 81).

[44] D. Cahalan. *American drinking practices*. 1969 (cit. on p. 81).

[45] K. E. Cairns, M. B. H. Yap, P. D. Pilkington, and A. F. Jorm. “Risk and protective factors for depression that adolescents can modify: a systematic review and meta-analysis of longitudinal studies”. In: *Journal of affective disorders* 169 (2014), pp. 61–75 (cit. on p. 92).

[46] *California School Climate, Health, and Learning Surveys*. `https://calschls.org/`. Accessed: 2021-05-08 (cit. on p. 82).

[47] Y. Cao, M. Long, J. Wang, H. Zhu, and Q. Wen. “Deep quantization network for efficient image retrieval”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016 (cit. on p. 45).

[48] N. Carlini and D. Wagner. “Audio adversarial examples: Targeted attacks on speech-to-text”. In: *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2018, pp. 1–7 (cit. on p. 18).

[49] N. Carlini and D. Wagner. “Defensive distillation is not robust to adversarial examples”. In: *arXiv preprint arXiv:1607.04311* (2016) (cit. on p. 20).

[50] N. Carlini and D. Wagner. “Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples”. In: *arXiv preprint arXiv:1711.08478* (2017) (cit. on p. 23).

[51] N. Carlini and D. Wagner. “Towards evaluating the robustness of neural networks”. In: *2017 ieee symposium on security and privacy (sp)*. IEEE. 2017, pp. 39–57 (cit. on p. 16).

[52] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht. “Adversarial Attacks for Tabular Data: Application to Fraud Detection and Imbalanced Data”. In: *arXiv preprint arXiv:2101.08030* (2021) (cit. on p. 77).

[53] J. Chang, J. Lee, A. Ha, et al. “Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples”. In: *Ophthalmology* 128.1 (2021), pp. 78–88 (cit. on p. 27).

[54] C. Chen, C. Qin, H. Qiu, et al. “Realistic adversarial data augmentation for MR image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 667–677 (cit. on p. 33).

[55] H. Chen, Q. Dou, L. Yu, J. Qin, and P. Heng. “VoxResNet: Deep voxelwise residual networks for braisegmentation from 3D MR images”. In: *NeuroImage* 170 (2018) (cit. on p. 43).

[56] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models”. In: *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 2017, pp. 15–26 (cit. on p. 77).

[57] S. Chen, W. Wang, and S. J. Pan. “Deep neural network quantization via layer-wise optimization using limited training data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 3329–3336 (cit. on p. 44).

[58] S. Chen, X. Huang, Z. He, and C. Sun. “DAmageNet: A Universal Adversarial Dataset”. In: *arXiv preprint arXiv:1912.07160* (2019) (cit. on p. 58).

[59] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. “Targeted backdoor attacks on deep learning systems using data poisoning”. In: *arXiv preprint arXiv:1712.05526* (2017) (cit. on p. 14).

[60] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. “A Survey of Model Compression and Acceleration for Deep Neural Networks”. In: arXiv/1710.09282 (2017) (cit. on p. 44).

[61] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, et al. “Opportunities and obstacles for deep learning in biology and medicine”. In: *Journal of The Royal Society Interface* 15.141 (2018), p. 20170387 (cit. on p. 4).

[62] K. Cho, B. Van Merriënboer, C. Gulcehre, et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014) (cit. on pp. 85, 87).

[63] Y. Choi, J. Choi, M. El-Khamy, and J. Lee. “Data-free network quantization with adversarial knowledge distillation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 710–711 (cit. on pp. 45, 46).

[64] F. Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258 (cit. on p. 57).

[65] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. “3D U-Net: Learning dense volumetric segmentation from sparse annotation”. In: *MICCAI*. Springer. 2016, pp. 424–432 (cit. on pp. 43, 48).

[66] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet. “Houdini: Fooling deep structured visual and speech recognition models with adversarial examples”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 18).

[67] J. P. Cohen, M. Luck, and S. Honari. “Distribution matching losses can hallucinate features in medical image translation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 529–536 (cit. on p. 61).

[68] J. K. Connor-Smith, B. E. Compas, M. E. Wadsworth, A. H. Thomsen, and H. Saltzman. “Responses to stress in adolescence: measurement of coping and involuntary stress responses.” In: *Journal of consulting and clinical psychology* 68.6 (2000), p. 976 (cit. on p. 81).

[69] F. Croce, M. Andriushchenko, V. Sehwag, et al. “RobustBench: a standardized adversarial robustness benchmark”. In: *arXiv preprint arXiv:2010.09670* (2020) (cit. on p. 58).

[70] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. “Autoaugment: Learning augmentation strategies from data”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 113–123 (cit. on p. 32).

[71] B. N. Cuthbert and T. R. Insel. “Toward the future of psychiatric diagnosis: the seven pillars of RDoC”. In: *BMC medicine* 11.1 (2013), pp. 1–8 (cit. on p. 79).

[72] M. A. Cyders, G. T. Smith, N. S. Spillane, S. Fischer, A. M. Annus, and C. Peterson. “Integration of impulsivity and positive mood to predict risky behavior: development and validation of a measure of positive urgency.” In: *Psychological assessment* 19.1 (2007), p. 107 (cit. on p. 81).

[73] Q. Dai, J. Wei, X. Shu, and Z. Feng. “Negativity bias for sad faces in depression: an event-related potential study”. In: *Clinical Neurophysiology* 127.12 (2016), pp. 3552–3560 (cit. on p. 79).

[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on pp. 16, 56, 57, 60).

[75] B. Desjardins, Y. Mirsky, M. P. Ortiz, et al. “DICOM images have been hacked! Now what?” In: *American Journal of Roentgenology* 214.4 (2020), pp. 727–735 (cit. on p. 5).

[76] T. DeVries and G. W. Taylor. “Dataset augmentation in feature space”. In: *arXiv preprint arXiv:1702.05538* (2017) (cit. on p. 32).

[77] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton, et al. “Stochastic activation pruning for robust adversarial defense”. In: *arXiv preprint arXiv:1803.01442* (2018) (cit. on p. 22).

[78] *Documentation for Binary Cross Entropy Loss in PyTorch 1.3.1 (BCEWithLogitsLoss)*. `https://pytorch.org/docs/1.3.1/nn.html#torch.nn.BCEWithLogitsLoss`. Accessed: 2021-05-09 (cit. on p. 88).

[79] K. Domschke, U. Dannlowski, C. Hohoff, et al. “Neuropeptide Y (NPY) gene: Impact on emotional processing and treatment response in anxious depression”. In: *European Neuropsychopharmacology* 20.5 (2010), pp. 301–309 (cit. on p. 79).

[80] C. Dwork, A. Roth, et al. “The algorithmic foundations of differential privacy.” In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), pp. 211–407 (cit. on p. 14).

[81] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994 (cit. on p. 86).

[82] M. Eichelberg, K. Kleber, and M. Kämmerer. “Cybersecurity in PACS and Medical Imaging: an Overview”. In: *Journal of Digital Imaging* (2020), pp. 1–16 (cit. on p. 5).

[83] L. F. and L. B. “Ternary Weight Networks”. In: *NIPS Workshop (EMDNN)* (2016) (cit. on p. 47).

[84] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard. “Adaptive data augmentation for image classification”. In: *2016 IEEE international conference on image processing (ICIP)*. Ieee. 2016, pp. 3688–3692 (cit. on p. 32).

[85] Y. Feng, B. Chen, T. Dai, and S.-T. Xia. “Adversarial attack on deep product quantization network for image retrieval”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 10786–10793 (cit. on p. 45).

[86] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. “Adversarial attacks on medical machine learning”. In: *Science* 363.6433 (2019), pp. 1287–1289 (cit. on pp. 5, 24).

[87] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox. “Adversarial examples for semantic image segmentation”. In: *arXiv preprint arXiv:1703.01101* (2017) (cit. on p. 18).

[88] R. A. Fisher. “The logic of inductive inference”. In: *Journal of the royal statistical society* 98.1 (1935), pp. 39–82 (cit. on p. 85).

[89] E. Fombonne, G. Wostear, V. Cooper, R. Harrington, and M. Rutter. “The Maudsley long-term follow-up of child and adolescent depression: I. Psychiatric outcomes in adulthood”. In: *The British Journal of Psychiatry* 179.3 (2001), pp. 210–217 (cit. on p. 78).

[90] E. E. Forbes and R. E. Dahl. “Research Review: altered reward function in adolescent depression: what, when and how?” In: *Journal of Child Psychology and Psychiatry* 53.1 (2012), pp. 3–15 (cit. on p. 92).

[91] S. S. Fredrick, M. K. Demaray, C. K. Malecki, and N. B. Dorio. “Can social support buffer the association between depression and suicidal ideation in adolescent boys and girls?” In: *Psychology in the Schools* 55.5 (2018), pp. 490–505 (cit. on p. 79).

[92] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. “Synthetic data augmentation using GAN for improved liver lesion classification”. In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 289–293 (cit. on pp. 33, 37).

[93] D. Fuhrmann, L. J. Knoll, and S.-J. Blakemore. “Adolescence as a sensitive period of brain development”. In: *Trends in cognitive sciences* 19.10 (2015), pp. 558–566 (cit. on p. 78).

[94] A. J. Fuligni and C. Hardway. “Daily variation in adolescents’ sleep, activities, and psychological well-being”. In: *Journal of Research on Adolescence* 16.3 (2006), pp. 353–378 (cit. on p. 91).

[95] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *arXiv preprint arXiv:1811.12231* (2018) (cit. on p. 57).

[96] M.-C. Geoffroy, M. Orri, A. Girard, L. C. Perret, and G. Turecki. “Trajectories of suicide attempts from early adolescence to emerging adulthood: Prospective 11-year follow-up of a Canadian cohort”. In: *Psychological medicine* (2020), pp. 1–11 (cit. on p. 78).

[97] J. N. Giedd, A. Raznahan, A. Alexander-Bloch, E. Schmitt, N. Gogtay, and J. L. Rapoport. “Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development”. In: *Neuropsychopharmacology* 40.1 (2015), pp. 43–49 (cit. on p. 78).

[98] G. A. Gioia, P. K. Isquith, S. C. Guy, and L. Kenworthy. *Behavior rating inventory of executive function: BRIEF*. Psychological Assessment Resources Odessa, FL, 2000 (cit. on p. 82).

[99] G. A. Gioia, P. K. Isquith, S. C. Guy, and L. Kenworthy. “Test review behavior rating inventory of executive function”. In: *Child Neuropsychology* 6.3 (2000), pp. 235–238 (cit. on p. 84).

[100] X. Glorot, A. Bordes, and Y. Bengio. “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323 (cit. on p. 13).

[101] P. I. Good. *Permutation, parametric, and bootstrap tests of hypotheses*. Springer Science & Business Media, 2006 (cit. on p. 86).

[102] I. J. Goodfellow, J. Shlens, and C. Szegedy. “Explaining and Harnessing Adversarial Examples”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015 (cit. on pp. 12–16, 19, 63).

[103] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. “Generative adversarial networks”. In: *arXiv preprint arXiv:1406.2661* (2014) (cit. on p. 32).

[104] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016 (cit. on pp. 3, 85, 87).

[105] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. “A very brief measure of the Big-Five personality domains”. In: *Journal of Research in personality* 37.6 (2003), pp. 504–528 (cit. on p. 81).

[106] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. “Badnets: Evaluating backdooring attacks on deep neural networks”. In: *IEEE Access* 7 (2019), pp. 47230–47244 (cit. on p. 14).

[107] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten. “Countering adversarial images using input transformations”. In: *arXiv preprint arXiv:1711.00117* (2017) (cit. on p. 21).

[108] Y. Guo, T. Ji, Q. Wang, L. Yu, and P. Li. “Quantized adversarial training: An iterative quantized local search approach”. In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2019, pp. 1066–1071 (cit. on p. 45).

[109] Y. Guo, C. Zhang, C. Zhang, and Y. Chen. “Sparse dnns with improved adversarial robustness”. In: *arXiv preprint arXiv:1810.09619* (2018) (cit. on pp. 45, 68).

[110] R. C. Gur, R. Sara, M. Hagendoorn, et al. “A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies”. In: *Journal of neuroscience methods* 115.2 (2002), pp. 137–143 (cit. on p. 83).

[111] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath. “Deep learning models for electrocardiograms are susceptible to adversarial attack”. In: *Nature medicine* 26.3 (2020), pp. 360–363 (cit. on p. 25).

[112] B. L. Hankin, L. Y. Abramson, T. E. Moffitt, P. A. Silva, R. McGee, and K. E. Angell. “Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study.” In: *Journal of abnormal psychology* 107.1 (1998), p. 128 (cit. on p. 92).

[113] M. Hashemi and A. Fathi. “PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards”. In: *arXiv preprint arXiv:2008.10138* (2020) (cit. on p. 78).

[114] B. P. Hasler, P. L. Franzen, M. de Zambotti, et al. “Eveningness and later sleep timing are associated with greater risk for alcohol and marijuana use in adolescence: initial findings from the national consortium on alcohol and neurodevelopment in adolescence study”. In: *Alcoholism: Clinical and Experimental Research* 41.6 (2017), pp. 1154–1165 (cit. on p. 81).

[115] G. Hasler, W. C. Drevets, H. K. Manji, and D. S. Charney. “Discovering endophenotypes for major depression”. In: *Neuropsychopharmacology* 29.10 (2004), pp. 1765–1781 (cit. on p. 79).

[116] T. Hazan, G. Papandreou, and D. Tarlow. *Perturbations, Optimization, and Statistics*. MIT Press, 2017 (cit. on pp. 12–15, 19).

[117] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 37, 57, 70).

[118] W. He, J. Wei, X. Chen, N. Carlini, and D. Song. “Adversarial example defense: Ensembles of weak defenses are not strong”. In: *11th {USENIX} workshop on offensive technologies ({WOOT} 17)*. 2017 (cit. on p. 22).

[119] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. “The digital database for screening mammography”. In: *Fifth International Workshop on Digital Mammography*. 2001, pp. 212–218 (cit. on pp. 37, 41).

[120] M. Heath, K. Bowyer, D. Kopans, et al. “Current status of the digital database for screening mammography”. In: *Digital mammography*. Springer, 1998, pp. 457–460 (cit. on p. 37).

[121] M. P. Heinrich, M. Blendowski, and O. Oktay. “TernaryNet: Faster deep model inference without GPUs for medical 3D segmentation using sparse and binary convolutions”. In: *IJCARS* 13.9 (2018), pp. 1311–1320 (cit. on pp. 44–47, 50).

[122] D. Hendrycks, S. Basart, N. Mu, et al. “The many faces of robustness: A critical analysis of out-of-distribution generalization”. In: *arXiv preprint arXiv:2006.16241* (2020) (cit. on pp. 4, 57).

[123] D. Hendrycks and T. Dietterich. “Benchmarking neural network robustness to common corruptions and perturbations”. In: *arXiv preprint arXiv:1903.12261* (2019) (cit. on pp. 56, 57).

[124] M. Hesselbrock, C. Easton, K. K. Bucholz, M. Schuckit, and V. Hesselbrock. "A validity study of the SSAGA-a comparison with the SCAN". In: *Addiction* 94.9 (1999), pp. 1361–1370 (cit. on p. 82).

[125] G. Hinton, O. Vinyals, and J. Dean. "Distilling the knowledge in a neural network". In: *NIPS Workshop* (2015) (cit. on pp. 20, 45, 50, 51).

[126] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 87).

[127] J. M. Holm-Denoma, B. L. Hankin, and J. F. Young. "Developmental trends of eating disorder symptoms and comorbid internalizing symptoms in children and adolescents". In: *Eating behaviors* 15.2 (2014), pp. 275–279 (cit. on p. 78).

[128] *How to steal modern NLP systems with gibberish?* `http://www.cleverhans.io/2020/04/06/stealing-bert.html`. Accessed: 2021-05-03 (cit. on pp. 3, 14).

[129] A. G. Howard, M. Zhu, B. Chen, et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017) (cit. on p. 65).

[130] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel. "Adversarial attacks on neural network policies". In: *arXiv preprint arXiv:1702.02284* (2017) (cit. on p. 19).

[131] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines". In: *NPJ digital medicine* 3.1 (2020), pp. 1–9 (cit. on pp. 76, 77).

[132] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. "Safety verification of deep neural networks". In: *International conference on computer aided verification*. Springer. 2017, pp. 3–29 (cit. on p. 60).

[133] T. R. Insel. "The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry". In: *American Journal of Psychiatry* 171.4 (2014), pp. 395–397 (cit. on p. 79).

[134] T. Insel, B. Cuthbert, M. Garvey, et al. *Research domain criteria (RDoC): toward a new classification framework for research on mental disorders*. 2010 (cit. on p. 79).

[135] *IoT Report: Imaging Systems Present Biggest Security Risk in Healthcare*. `https://www.hcinnovationgroup.com/cybersecurity/news/13029895/iot-report-imaging-systems-present-biggest-security-risk-in-healthcare` (cit. on p. 4).

[136] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 11–19 (cit. on p. 66).

[137] F.-H. Jhang. "Uncontrollable and controllable negative life events and changes in mental health problems: Exploring the moderation effects of family support and self-efficacy in economically disadvantaged adolescents". In: *Children and Youth Services Review* 118 (2020), p. 105417 (cit. on p. 91).

[138] L. Jiang, K. Qiao, R. Qin, et al. "Cycle-consistent adversarial GAN: The integration of adversarial attack and defense". In: *Security and Communication Networks* 2020 (2020) (cit. on p. 17).

[139] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang. "Ape-gan: Adversarial perturbation elimination with gan". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3842–3846 (cit. on p. 23).

[140] P. K.M, S. EV, P. S., et al. *The NCANDA_PUBLIC_6Y_REDCAP_V01 Data Release of the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA)*. 2021 (cit. on p. 85).

[141] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren. "Secure, privacy-preserving and federated machine learning in medical imaging". In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311 (cit. on p. 14).

[142] C. Kamann and C. Rother. “Benchmarking the robustness of semantic segmentation models with respect to common corruptions”. In: *International Journal of Computer Vision* 129.2 (2021), pp. 462–483 (cit. on p. 57).

[143] C. Kanbak, S.-M. Moosavi-Dezfooli, and P. Frossard. “Geometric robustness of deep networks: analysis and improvement”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4441–4449 (cit. on pp. 17, 30, 33–35).

[144] D. Kang, Y. Sun, D. Hendrycks, T. Brown, and J. Steinhardt. “Testing robustness against unforeseen adversaries”. In: *arXiv preprint arXiv:1908.08016* (2019) (cit. on p. 58).

[145] H. Kannan, A. Kurakin, and I. Goodfellow. “Adversarial logit pairing”. In: *arXiv preprint arXiv:1803.06373* (2018) (cit. on p. 20).

[146] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. “Reluplex: An efficient SMT solver for verifying deep neural networks”. In: *International Conference on Computer Aided Verification*. Springer. 2017, pp. 97–117 (cit. on pp. 45, 60).

[147] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. “Key challenges for delivering clinical impact with artificial intelligence”. In: *BMC medicine* 17.1 (2019), pp. 1–9 (cit. on p. 31).

[148] K. S. Kendler, C. M. Bulik, J. Silberg, J. M. Hettema, J. Myers, and C. A. Prescott. “Childhood sexual abuse and adult psychiatric and substance use disorders in women: an epidemiological and cotwin control analysis”. In: *Archives of general psychiatry* 57.10 (2000), pp. 953–959 (cit. on p. 90).

[149] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters. “Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication”. In: *Archives of general psychiatry* 62.6 (2005), pp. 593–602 (cit. on p. 78).

[150] A. Khakzar, S. Albarqouni, and N. Navab. “Learning Interpretable Features via Adversarially Robust Optimization”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 793–800 (cit. on p. 27).

[151] F. Khalid, H. Ali, H. Tariq, et al. “Qusecnets: Quantization-based defense mechanism for securing deep neural network against adversarial attacks”. In: *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*. IEEE. 2019, pp. 182–187 (cit. on p. 45).

[152] S. T. Kim, L. Goli, M. Paschali, et al. “Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs”. In: *arXiv preprint arXiv:2103.07240* (2021) (cit. on pp. 76, 77).

[153] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 49).

[154] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 66, 88).

[155] W. K. Kirchner. “Age differences in short-term retention of rapidly changing information.” In: *Journal of experimental psychology* 55.4 (1958), p. 352 (cit. on p. 83).

[156] S. Kitamura, A. Hida, M. Watanabe, et al. “Evening preference is related to the incidence of depressive states independent of sleep-wake conditions”. In: *Chronobiology international* 27.9-10 (2010), pp. 1797–1812 (cit. on p. 91).

[157] J. Klinger-Koenig, J. Hertel, J. Terock, H. Voelzke, S. Van der Auwera, and H. J. Grabe. “Predicting physical and mental health symptoms: Additive and interactive effects of difficulty identifying feelings, neuroticism and extraversion”. In: *Journal of psychosomatic research* 115 (2018), pp. 14–23 (cit. on pp. 79, 90).

[158] *Know your enemy. How you can create and defend against adversarial attacks.* `https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3`. Accessed: 2021-05-03 (cit. on p. 16).

[159] P. W. Koh, S. Sagawa, H. Marklund, et al. “Wilds: A benchmark of in-the-wild distribution shifts”. In: *arXiv preprint arXiv:2012.07421* (2020) (cit. on pp. 57, 58).

[160] E. Kokiopoulou and P. Frossard. “Minimum distance between pattern transformation manifolds: Algorithm and applications”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.7 (2009), pp. 1225–1238 (cit. on p. 35).

[161] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. “Federated Learning: Strategies for improving communication efficiency”. In: *NIPS Workshop* (2016) (cit. on p. 44).

[162] K. Krishna, G. S. Tomar, A. P. Parikh, N. Papernot, and M. Iyyer. “Thieves on sesame street! model extraction of bert-based apis”. In: *arXiv preprint arXiv:1910.12366* (2019) (cit. on p. 14).

[163] A. Kurakin, I. J. Goodfellow, and S. Bengio. “Adversarial Machine Learning at Scale”. In: *CoRR* abs/1611.01236 (2016) (cit. on p. 19).

[164] A. Kurakin, I. Goodfellow, S. Bengio, et al. *Adversarial examples in the physical world*. 2016 (cit. on p. 16).

[165] Z. A. Kyte, I. M. Goodyer, and B. J. Sahakian. “Selected executive skills in adolescents with recent first episode major depression”. In: *Journal of Child Psychology and Psychiatry* 46.9 (2005), pp. 995–1005 (cit. on p. 91).

[166] B. Landman and S. Warfield. “MICCAI workshop on Multiatlas labeling”. In: *MICCAI Grand Challenge* (2012) (cit. on p. 66).

[167] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on pp. 16, 24).

[168] C. Lee, J. Yoon, and M. Van Der Schaar. “Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data”. In: *IEEE Transactions on Biomedical Engineering* 67.1 (2019), pp. 122–133 (cit. on p. 76).

[169] J. LeMoult, K. L. Humphreys, A. Tracy, J.-A. Hoffmeister, E. Ip, and I. H. Gotlib. “Meta-analysis: exposure to early life stress and risk for depression in childhood and adolescence”. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 59.7 (2020), pp. 842–855 (cit. on p. 78).

[170] A. Leslie, A. Jones, and P. Goddard. “The influence of clinical information on the reporting of CT by radiologists.” In: *The British journal of radiology* 73.874 (2000), pp. 1052–1055 (cit. on p. 76).

[171] E. Levy, Y. Mathov, Z. Katzir, A. Shabtai, and Y. Elovici. “Not All Datasets Are Born Equal: On Heterogeneous Data and Adversarial Examples”. In: *arXiv preprint arXiv:2010.03180* (2020) (cit. on p. 78).

[172] B. Li, C. Chen, W. Wang, and L. Carin. “Certified adversarial robustness with additive noise”. In: *arXiv preprint arXiv:1809.03113* (2018) (cit. on p. 21).

[173] X. Li and D. Zhu. “Robust detection of adversarial attacks on medical images”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1154–1158 (cit. on p. 26).

[174] Y. Li, H. Zhang, C. Bermudez, Y. Chen, B. A. Landman, and Y. Vorobeychik. “Anatomical context protects deep learning from adversarial perturbations in medical imaging”. In: *Neurocomputing* 379 (2020), pp. 370–378 (cit. on pp. 25, 75, 78).

[175] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu. “Defense against adversarial attacks using high-level representation guided denoiser”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1778–1787 (cit. on p. 23).

[176] J. Lin, C. Gan, and S. Han. “Defensive Quantization: When Efficiency Meets Robustness”. In: *ICLR*. 2019 (cit. on pp. 44, 45, 68).

[177] S. Liu, A. A. A. Setio, F. C. Ghesu, et al. “No surprises: Training robust lung nodule detection for low-dose CT scans by augmenting with adversarial attacks”. In: *IEEE Transactions on Medical Imaging* 40.1 (2020), pp. 335–345 (cit. on p. 26).

[178] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh. “Towards robust neural networks via random self-ensemble”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 369–385 (cit. on p. 21).

[179] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk. “Improving robustness without sacrificing accuracy with patch gaussian augmentation”. In: *arXiv preprint arXiv:1906.02611* (2019) (cit. on p. 32).

[180] N. Lovato and M. Gradisar. “A meta-analysis and model of the relationship between sleep and depression in adolescents: recommendations for future research and clinical practice”. In: *Sleep medicine reviews* 18.6 (2014), pp. 521–529 (cit. on p. 79).

[181] T. Luo, T. Cai, M. Zhang, S. Chen, and L. Wang. “RANDOM MASK: Towards Robust Convolutional Neural Networks”. In: *arXiv preprint arXiv:2007.14249* (2020) (cit. on p. 22).

[182] X. Ma, Y. Niu, L. Gu, et al. “Understanding adversarial attacks on deep learning based medical image analysis systems”. In: *Pattern Recognition* 110 (2021), p. 107332 (cit. on p. 25).

[183] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017) (cit. on pp. 16, 63).

[184] H. B. Mann and D. R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60 (cit. on p. 87).

[185] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray. “Metric learning for adversarial robustness”. In: *arXiv preprint arXiv:1909.00900* (2019) (cit. on p. 23).

[186] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. “Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults”. In: *J. Cognitive Neuroscience* 19.9 (2007), pp. 1498–1507 (cit. on pp. 18, 48, 50, 66).

[187] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. “Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults”. In: *Journal of cognitive neuroscience* 19.9 (2007), pp. 1498–1507 (cit. on p. 66).

[188] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi. “Bagan: Data augmentation with balancing gan”. In: *arXiv preprint arXiv:1803.09655* (2018) (cit. on p. 32).

[189] N. R. Marmorstein. “Longitudinal associations between alcohol problems and depressive symptoms: early adolescence through early adulthood”. In: *Alcoholism: Clinical and Experimental Research* 33.1 (2009), pp. 49–59 (cit. on p. 92).

[190] G. Marum, J. Clench-Aas, R. B. Nes, and R. K. Raanaas. “The relationship between negative life events, psychological distress and life satisfaction: a population-based study”. In: *Quality of Life Research* 23.2 (2014), pp. 601–611 (cit. on p. 91).

[191] K. Matthews, D. Coghill, and S. Rhodes. “Neuropsychological functioning in depressed adolescent girls”. In: *Journal of Affective Disorders* 111.1 (2008), pp. 113–118 (cit. on p. 79).

[192] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir. “Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI”. In: *Journal of Magnetic Resonance Imaging* 49.4 (2019), pp. 939–954 (cit. on p. 43).

[193] K. A. McLaughlin, J. G. Green, M. J. Gruber, N. A. Sampson, A. M. Zaslavsky, and R. C. Kessler. “Childhood adversities and first onset of psychiatric disorders in a national sample of US adolescents”. In: *Archives of general psychiatry* 69.11 (2012), pp. 1151–1160 (cit. on p. 79).

[194] G. C. Medeiros, A. J. Rush, M. Jha, et al. "Positive and negative valence systems in major depression have distinct clinical features, response to antidepressants, and relationships with immunomarkers". In: *Depression and anxiety* 37.8 (2020), pp. 771–783 (cit. on p. 79).

[195] D. Meng and H. Chen. "Magnet: a two-pronged defense against adversarial examples". In: *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017, pp. 135–147 (cit. on p. 23).

[196] K. R. Merikangas, J.-p. He, M. Burstein, et al. "Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A)". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 49.10 (2010), pp. 980–989 (cit. on p. 78).

[197] A. D. Meruelo, T. Brumback, B. J. Nagel, F. C. Baker, S. A. Brown, and S. F. Tapert. "Neuroimaging markers of adolescent depression in the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) study". In: *Journal of Affective Disorders* (2021) (cit. on p. 92).

[198] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571 (cit. on pp. 43, 48).

[199] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici. "CT-GAN: Malicious tampering of 3D medical imagery using deep learning". In: *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 2019, pp. 461–478 (cit. on pp. 4, 5).

[200] T. H. Monk, D. J. Buysse, K. S. Kennedy, J. M. Potts, J. M. DeGrazia, and J. M. Miewald. "Measuring sleep habits without using a diary: the sleep timing questionnaire". In: *Sleep* 26.2 (2003), pp. 208–212 (cit. on p. 81).

[201] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. "Universal adversarial perturbations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1765–1773 (cit. on p. 16).

[202] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. "Deepfool: a simple and accurate method to fool deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2574–2582 (cit. on pp. 16, 63).

[203] *More Than 45 Million Medical Images Openly Accessible Online*.
https://cybelangel.com/blog/medical-data-leaks/ (cit. on p. 5).

[204] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012 (cit. on pp. 87, 88).

[205] J. Nalepa, G. Mrukwa, S. Piechaczek, et al. "Data augmentation via image registration". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 4250–4254 (cit. on p. 32).

[206] N. Ng, K. Cho, and M. Ghassemi. "Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness". In: *arXiv preprint arXiv:2009.10195* (2020) (cit. on p. 32).

[207] *NTT Security, 2017 Global Threat Intelligence Report (GTIR)*.
https://us.nttdata.com/en/-/media/nttdataamerica/files/americasd2/infrastructure_managed_services/gtir-ntt-security-ntt-data-04252017.pdf. Accessed: 2021-05-06 (cit. on p. 5).

[208] E. Okafor, L. Schomaker, and M. A. Wiering. "An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals". In: *Journal of Information and Telecommunication* 2.4 (2018), pp. 465–491 (cit. on p. 32).

[209] G. Ortiz-Jiménez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard. "Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness". In: *Proceedings of the IEEE* (2021) (cit. on p. 19).

[210] N. Papernot and P. McDaniel. “Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning”. In: *arXiv preprint arXiv:1803.04765* (2018) (cit. on p. 23).

[211] N. Papernot and P. D. McDaniel. “Extending Defensive Distillation”. In: *CoRR* abs/1705.05264 (2017) (cit. on p. 20).

[212] N. Papernot, P. McDaniel, and I. Goodfellow. “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples”. In: *arXiv preprint arXiv:1605.07277* (2016) (cit. on pp. 11, 12).

[213] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. “The limitations of deep learning in adversarial settings”. In: *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE. 2016, pp. 372–387 (cit. on pp. 12, 15, 16, 63, 78).

[214] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. “Distillation as a defense to adversarial perturbations against deep neural networks”. In: *2016 IEEE symposium on security and privacy (SP)*. IEEE. 2016, pp. 582–597 (cit. on p. 20).

[215] M. Paschali, S. Conjeti, F. Navarro, and N. Navab. “Generalizability vs. robustness: investigating medical imaging networks using adversarial examples”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 493–501 (cit. on pp. 4, 18, 24, 41, 55, 62, 65, 66, 68, 69).

[216] M. Paschali, S. Gasperini, A. G. Roy, M. Y.-S. Fang, and N. Navab. “3dq: Compact quantized neural networks for volumetric whole brain segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 438–446 (cit. on pp. 4, 43, 47–50, 52).

[217] M. Paschali, W. Simson, A. G. Roy, R. Göbl, C. Wachinger, and N. Navab. “Manifold exploring data augmentation with geometric transformations for increased performance and robustness”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 517–529 (cit. on pp. 4, 17, 29, 30, 36, 38–41).

[218] A. Paszke, S. Gross, F. Massa, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035 (cit. on pp. 37, 49, 88).

[219] K. Pei, Y. Cao, J. Yang, and S. Jana. “Deepxplore: Automated whitebox testing of deep learning systems”. In: *proceedings of the 26th Symposium on Operating Systems Principles*. 2017, pp. 1–18 (cit. on p. 57).

[220] L. S. Penrose and R. Penrose. “Impossible objects: A special type of visual illusion”. In: *British Journal of Psychology* 49.1 (1958), pp. 31–33 (cit. on p. 12).

[221] *Poisoning attacks on Machine Learning*. `https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db`. Accessed: 2021-05-03 (cit. on p. 14).

[222] R. J. Porter, P. Gallagher, J. M. Thompson, and A. H. Young. “Neurocognitive impairment in drug-free patients with major depressive disorder”. In: *The British Journal of Psychiatry* 182.3 (2003), pp. 214–220 (cit. on p. 79).

[223] S. Pouyanfar, S. Sadiq, Y. Yan, et al. “A survey on deep learning: Algorithms, techniques, and applications”. In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–36 (cit. on p. 3).

[224] L. Pulina and A. Tacchella. “An abstraction-refinement approach to verification of artificial neural networks”. In: *International Conference on Computer Aided Verification*. Springer. 2010, pp. 243–257 (cit. on p. 60).

[225] A. Radford, L. Metz, and S. Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv preprint arXiv:1511.06434* (2015) (cit. on p. 39).

[226] A. Raghunathan, J. Steinhardt, and P. Liang. “Certified defenses against adversarial examples”. In: *arXiv preprint arXiv:1801.09344* (2018) (cit. on p. 24).

[227] A. Raghunathan, J. Steinhardt, and P. Liang. “Semidefinite relaxations for certifying robustness to adversarial examples”. In: *arXiv preprint arXiv:1811.01057* (2018) (cit. on p. 24).

[228] A. Rahmati, S.-M. Moosavi-Dezfooli, P. Frossard, and H. Dai. “GeoDA: a geometric framework for black-box adversarial attacks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8446–8455 (cit. on p. 17).

[229] A. Rakhsha, G. Radanovic, R. Devidze, X. Zhu, and A. Singla. “Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7974–7984 (cit. on p. 19).

[230] D. Ramachandram and G. W. Taylor. “Deep multimodal learning: A survey on recent advances and trends”. In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108 (cit. on p. 76).

[231] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. “XNOR-Net: Imagenet classification using binary convolutional neural networks”. In: *ECCV*. Springer. 2016, pp. 525–542 (cit. on pp. 44, 47).

[232] J. Rauber, W. Brendel, and M. Bethge. “Foolbox v0. 8.0: A Python toolbox to benchmark the robustness of machine learning models. CoRR abs/1707.04131 (2017)”. In: *arXiv preprint arXiv:1707.04131* (2017) (cit. on p. 67).

[233] K. Ren, T. Zheng, Z. Qin, and X. Liu. “Adversarial attacks and defenses in deep learning”. In: *Engineering* 6.3 (2020), pp. 346–360 (cit. on p. 45).

[234] J. P. Rice, T. Reich, K. K. Bucholz, et al. “Comparison of direct interview and family history diagnoses of alcohol dependence”. In: *Alcoholism: Clinical and Experimental Research* 19.4 (1995), pp. 1018–1023 (cit. on p. 82).

[235] T. Roenneberg, T. Kuehnle, P. P. Pramstaller, et al. “A marker for the end of adolescence”. In: *Current biology* 14.24 (2004), R1038–R1039 (cit. on p. 91).

[236] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241 (cit. on pp. 23, 50, 62, 66).

[237] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger. “QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy”. In: *NeuroImage* 186 (2019), pp. 713–727 (cit. on pp. 43, 48).

[238] A. G. Roy, S. Conjeti, D. Sheet, A. Katouzian, N. Navab, and C. Wachinger. “Error corrective boosting for learning fully convolutional networks with limited data”. In: *MICCAI*. Springer. 2017, pp. 231–239 (cit. on pp. 37, 49, 65, 66).

[239] W. J. Rudman, J. S. Eberhardt, W. Pierce, and S. Hart-Hester. “Healthcare fraud and abuse”. In: *Perspectives in Health Information Management/AHIMA, American Health Information Management Association* 6.Fall (2009) (cit. on pp. 4, 5).

[240] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet. “Adversarial Manipulation of Deep Representations”. In: *CoRR* abs/1511.05122 (2015) (cit. on p. 19).

[241] P. Samangouei, M. Kabkab, and R. Chellappa. “Defense-gan: Protecting classifiers against adversarial attacks using generative models”. In: *arXiv preprint arXiv:1805.06605* (2018) (cit. on p. 22).

[242] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers. “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks”. In: *Scientific reports* 9.1 (2019), pp. 1–9 (cit. on p. 33).

[243] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520 (cit. on p. 57).

[244] C. A. Sanislow, D. S. Pine, K. J. Quinn, et al. “Developing constructs for psychopathology research: research domain criteria.” In: *Journal of abnormal psychology* 119.4 (2010), p. 631 (cit. on p. 80).

[245] S. Santurkar, D. Tsipras, and A. Madry. “Breeds: Benchmarks for subpopulation shift”. In: *arXiv preprint arXiv:2008.04859* (2020) (cit. on p. 58).

[246] T. Schulte, J.-Y. Hong, E. V. Sullivan, et al. “Effects of age, sex, and puberty on neural efficiency of cognitive and motor control in adolescents”. In: *Brain imaging and behavior* (2019), pp. 1–19 (cit. on p. 83).

[247] A. Shafahi, W. R. Huang, M. Najibi, et al. “Poison frogs! targeted clean-label poisoning attacks on neural networks”. In: *arXiv preprint arXiv:1804.00792* (2018) (cit. on p. 14).

[248] A. Shafahi, M. Najibi, A. Ghiasi, et al. “Adversarial training for free!” In: *arXiv preprint arXiv:1904.12843* (2019) (cit. on p. 20).

[249] A. Shafahi, M. Najibi, Z. Xu, J. Dickerson, L. S. Davis, and T. Goldstein. “Universal adversarial training”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5636–5643 (cit. on p. 20).

[250] Y. Sharma and P.-Y. Chen. “Bypassing feature squeezing by increasing adversary strength”. In: *arXiv preprint arXiv:1803.09868* (2018) (cit. on p. 22).

[251] R. Shaw, C. Sudre, S. Ourselin, and M. J. Cardoso. “MRI k-space motion artefact augmentation: model robustness and task-specific uncertainty”. In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2019, pp. 427–436 (cit. on p. 61).

[252] R. S. Shim, P. Baltrus, J. Ye, and G. Rust. “Prevalence, treatment, and control of depressive symptoms in the United States: results from the National Health and Nutrition Examination Survey (NHANES), 2005–2008”. In: *The Journal of the American Board of Family Medicine* 24.1 (2011), pp. 33–38 (cit. on p. 79).

[253] M. A. Short, S. A. Booth, O. Omar, L. Ostlundh, and T. Arora. “The relationship between sleep duration and mood in adolescents: A systematic review and meta-analysis”. In: *Sleep medicine reviews* 52 (2020), p. 101311 (cit. on pp. 78, 79).

[254] C. Shorten and T. M. Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of Big Data* 6.1 (2019), pp. 1–48 (cit. on p. 31).

[255] K. Simonyan, A. Vedaldi, and A. Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013) (cit. on p. 16).

[256] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on p. 37).

[257] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, et al. “A large annotated medical image dataset for the development and evaluation of segmentation algorithms”. In: *arXiv preprint arXiv:1902.09063* (2019) (cit. on p. 48).

[258] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. “Certifying some distributional robustness with principled adversarial training”. In: *arXiv preprint arXiv:1710.10571* (2017) (cit. on p. 24).

[259] C. S. Smith, C. Reilly, and K. Midkiff. “Evaluation of three circadian rhythm questionnaires with suggestions for an improved measure of morningness.” In: *Journal of Applied psychology* 74.5 (1989), p. 728 (cit. on p. 81).

[260] C. Song, E. Fallon, and H. Li. “Improving Adversarial Robustness in Weight-quantized Neural Networks”. In: *arXiv preprint arXiv:2012.14965* (2020) (cit. on p. 45).

[261] C. Spearman. “The proof and measurement of association between two things”. In: *The American journal of psychology* 15 (1904), pp. 72–101 (cit. on p. 87).

[262] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. “Striving for simplicity: The all convolutional net”. In: *arXiv preprint arXiv:1412.6806* (2014) (cit. on p. 86).

[263] C. Stanger, S. R. Ryan, H. Fu, et al. “Delay discounting predicts adolescent substance abuse treatment outcome.” In: *Experimental and clinical psychopharmacology* 20.3 (2012), p. 205 (cit. on p. 83).

[264] Y. Stikkelbroek, D. H. Bodden, M. Kleinjan, M. Reijnders, and A. L. van Baar. “Adolescent depression and negative life events, the mediating role of cognitive emotion regulation”. In: *PloS one* 11.8 (2016), e0161062 (cit. on p. 90).

[265] M. Stites and O. S. Pianykh. “How secure is your radiology department? Mapping digital radiology adoption and security worldwide”. In: *American Journal of Roentgenology* 206.4 (2016), pp. 797–804 (cit. on p. 5).

[266] J. R. Stroop. “Studies of interference in serial verbal reactions.” In: *Journal of experimental psychology* 18.6 (1935), p. 643 (cit. on p. 83).

[267] J. Su, D. V. Vargas, and K. Sakurai. “One pixel attack for fooling deep neural networks”. In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 828–841 (cit. on p. 17).

[268] E. V. Sullivan, T. Brumback, S. F. Tapert, et al. “Cognitive, emotion control, and motor performance of adolescents in the NCANDA study: Contributions from alcohol consumption, age, sex, ethnicity, and family history of addiction.” In: *Neuropsychology* 30.4 (2016), p. 449 (cit. on p. 83).

[269] T. Suslow, U. Dannlowski, J. Lalee-Mentzel, U.-S. Donges, V. Arolt, and A. Kersting. “Spatial processing of facial emotion in patients with unipolar depression: a longitudinal study”. In: *Journal of affective disorders* 83.1 (2004), pp. 59–63 (cit. on p. 91).

[270] J. R. Swartz and C. S. Monk. “The role of corticolimbic circuitry in the development of anxiety disorders in children and adolescents”. In: *The neurobiology of childhood* (2013), pp. 133–148 (cit. on p. 78).

[271] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. on p. 37).

[272] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826 (cit. on pp. 62, 65).

[273] C. Szegedy, W. Zaremba, I. Sutskever, et al. “Intriguing properties of neural networks”. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2014 (cit. on pp. 3, 11–13, 15, 36, 61, 63).

[274] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. “When Robustness Doesn’t Promote Robustness: Synthetic vs. Natural Distribution Shifts on ImageNet”. In: (2019) (cit. on p. 57).

[275] E. H. Telzer, D. Goldenberg, A. J. Fuligni, M. D. Lieberman, and A. Gálvan. “Sleep variability in adolescence is associated with altered brain development”. In: *Developmental cognitive neuroscience* 14 (2015), pp. 16–22 (cit. on p. 92).

[276] S. Thaler and V. Menkovski. “The Role of Deep Learning in Improving Healthcare”. In: *Data Science for Healthcare - Methodologies and Applications*. 2019, pp. 75–116 (cit. on p. 44).

[277] *The challenge of verification and testing of machine learning*. http://www.cleverhans.io/security/privacy/ml/2017/06/14/verification.html. Accessed: 2021-05-03 (cit. on pp. 59, 60).

[278] P. A. Thoits. "Mechanisms linking social ties and support to physical and mental health". In: *Journal of health and social behavior* 52.2 (2011), pp. 145–161 (cit. on p. 79).

[279] Y. Tian, K. Pei, S. Jana, and B. Ray. "Deeptest: Automated testing of deep-neural-network-driven autonomous cars". In: *Proceedings of the 40th international conference on software engineering*. 2018, pp. 303–314 (cit. on p. 57).

[280] M. Tirindelli, C. Eilers, W. Simson, M. Paschali, M. Farid Azampour, and N. Navab. "Rethinking Ultrasound Augmentation: A Physics-Inspired Approach". In: *arXiv preprint arXiv:2105.02188* (2021) (cit. on p. 33).

[281] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. D. McDaniel. "Ensemble Adversarial Training: Attacks and Defenses". In: *CoRR* abs/1705.07204 (2017) (cit. on p. 19).

[282] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. "The space of transferable adversarial examples". In: *arXiv preprint arXiv:1704.03453* (2017) (cit. on pp. 12, 13, 69).

[283] M. T. Treadway and D. H. Zald. "Reconsidering anhedonia in depression: lessons from translational neuroscience". In: *Neuroscience & Biobehavioral Reviews* 35.3 (2011), pp. 537–555 (cit. on p. 79).

[284] P. Tschandl, C. Rosendahl, and H. Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5.1 (2018), pp. 1–9 (cit. on p. 40).

[285] L. W. Tu. *Differential geometry: connections, curvature, and characteristic classes*. Vol. 275. Springer, 2017 (cit. on p. 35).

[286] D. V. Vargas and J. Su. "Understanding the one-pixel attack: Propagation maps and locality analysis". In: *arXiv preprint arXiv:1902.02947* (2019) (cit. on p. 17).

[287] R. Vivanti, L. Joskowicz, N. Lev-Cohain, A. Ephrat, and J. Sosna. "Patient-specific and global convolutional neural networks for robust automatic liver tumor delineation in follow-up CT studies". In: *MBEC* 56.9 (2018), pp. 1699–1713 (cit. on p. 44).

[288] F. Wang, L. P. Casalino, and D. Khullar. "Deep learning in medicine—promise, progress, and challenges". In: *JAMA internal medicine* 179.3 (2019), pp. 293–294 (cit. on p. 4).

[289] J. Wang, L. Perez, et al. "The effectiveness of data augmentation in image classification using deep learning". In: *Convolutional Neural Networks Vis. Recognit* 11 (2017) (cit. on p. 32).

[290] D. Watson, S. M. Stasik, S. Ellickson-Larew, and K. Stanton. "Extraversion and psychopathology: A facet-level analysis." In: *Journal of abnormal psychology* 124.2 (2015), p. 432 (cit. on p. 90).

[291] T.-W. Weng, H. Zhang, P.-Y. Chen, et al. "Evaluating the robustness of neural networks: An extreme value theory approach". In: *arXiv preprint arXiv:1801.10578* (2018) (cit. on p. 60).

[292] S. C. Wetstein, C. González-Gonzalo, G. Bortsova, et al. "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors". In: *arXiv preprint arXiv:2006.06356* (2020) (cit. on p. 26).

[293] H. C. Wilcox and J. C. Anthony. "Child and adolescent clinical features as forerunners of adult-onset major depressive disorder: retrospective evidence from an epidemiological sample". In: *Journal of affective disorders* 82.1 (2004), pp. 9–20 (cit. on p. 79).

[294] E. Wong and Z. Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5286–5295 (cit. on p. 24).

[295] E. Wong, F. R. Schmidt, J. H. Metzen, and J. Z. Kolter. “Scaling provable adversarial defenses”. In: *arXiv preprint arXiv:1805.12514* (2018) (cit. on p. 24).

[296] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. “Understanding data augmentation for classification: when to warp?” In: *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE. 2016, pp. 1–6 (cit. on p. 32).

[297] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song. “Generating adversarial examples with adversarial networks”. In: *arXiv preprint arXiv:1801.02610* (2018) (cit. on p. 17).

[298] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. “Spatially transformed adversarial examples”. In: *arXiv preprint arXiv:1801.02612* (2018) (cit. on p. 17).

[299] K. Y. Xiao, V. Tjeng, N. M. Shafiullah, and A. Madry. “Training for faster adversarial robustness verification via inducing relu stability”. In: *arXiv preprint arXiv:1809.03008* (2018) (cit. on p. 45).

[300] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. “Mitigating adversarial effects through randomization”. In: *arXiv preprint arXiv:1711.01991* (2017) (cit. on pp. 20, 21).

[301] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. “Adversarial examples for semantic segmentation and object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1369–1378 (cit. on pp. 18, 63).

[302] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang. “Towards evaluating the robustness of deep diagnostic models by adversarial attack”. In: *Medical Image Analysis* 69 (2021), p. 101977 (cit. on p. 26).

[303] W. Xu, D. Evans, and Y. Qi. “Feature squeezing mitigates and detects carlini/wagner adversarial examples”. In: *arXiv preprint arXiv:1705.10686* (2017) (cit. on p. 22).

[304] W. Xu, D. Evans, and Y. Qi. “Feature squeezing: Detecting adversarial examples in deep neural networks”. In: *arXiv preprint arXiv:1704.01155* (2017) (cit. on p. 22).

[305] F.-F. Xue, J. Peng, R. Wang, Q. Zhang, and W.-S. Zheng. “Improving robustness of medical image diagnosis with denoising convolutional neural networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 846–854 (cit. on p. 26).

[306] Q. Yao, Z. He, H. Han, and S. K. Zhou. “Miss the Point: Targeted Adversarial Attack on Multiple Landmark Detection”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 692–702 (cit. on p. 25).

[307] X. Ying. “An overview of overfitting and its solutions”. In: *Journal of Physics: Conference Series*. Vol. 1168. 2. IOP Publishing. 2019, p. 022022 (cit. on pp. 4, 88).

[308] W. Yoon, J. Park, and D. Kim. “Stochastic quantized activation: To prevent overfitting in fast adversarial training”. In: (2018) (cit. on p. 45).

[309] M. de Zambotti, A. Goldstone, I. M. Colrain, and F. C. Baker. “Insomnia disorder in adolescence: diagnosis, impact, and treatment”. In: *Sleep medicine reviews* 39 (2018), pp. 12–24 (cit. on p. 78).

[310] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca. “Data augmentation using learned transformations for one-shot medical image segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8543–8553 (cit. on p. 33).

[311] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. “Random erasing data augmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 13001–13008 (cit. on p. 37).

[312] C. Zhu, S. Han, H. Mao, and W. J. Dally. “Trained ternary quantization”. In: *ICLR* (2017) (cit. on pp. 44, 46, 47, 49).

[313] 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*. 2017, pp. 2223–2232 (cit. on p. 33).

[314] W. Zhu, X. Xiang, T. D. Tran, G. D. Hager, and X. Xie. “Adversarial deep structured nets for mass segmentation from mammograms”. In: *15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018*. IEEE, 2018, pp. 847–850 (cit. on p. 61).

[315] D. Zimmerer, J. Petersen, G. Köhler, et al. *Medical Out-of-Distribution Analysis Challenge 2021*. Mar. 2021 (cit. on p. 59).

[316] D. Zügner and S. Günnemann. “Adversarial attacks on graph neural networks via meta learning”. In: *arXiv preprint arXiv:1902.08412* (2019) (cit. on p. 19).

List of Figures

List of Tables