FernUniversität in Hagen
Faculty of Mathematics and Computer Science

AIG
Artificial Intelligence Group

Leveraging Synthetic Datasets for Enhanced Optimization of Mask R-CNNs through Comparative Analysis

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Praktische Informatik

submitted by
Tobias Wolf

First examiner:	Prof. Dr. Matthias Thimm Artificial Intelligence Group
Advisor:	Dr. Marvin Hoffmann FUCHS LUBRICANTS GERMANY GmbH

Statement

Ich erkläre, dass ich die Masterarbeit selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe. Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht. Die Versicherung selbstständiger Arbeit gilt auch für enthaltene Zeichnungen, Skizzen oder graphische Darstellungen. Die Arbeit wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Mit der Abgabe der elektronischen Fassung der endgültigen Version der Arbeit nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate geprüft werden kann und ausschließlich für Prüfungszwecke gespeichert wird.

	Yes	No
I agree to have this thesis published in the library.	☑	☐
I agree to have this thesis published on the webpage of the artificial intelligence group.	☑	☐
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	☑	☐
The source code is available under a GNU General Public License (GPLv3).	☑	☐
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	☑	☐

Mannheim, 27.05.2024

(Place, Date) (Signature)

Zusammenfassung

Die Verfügbarkeit ausreichender und vielfältiger Datensätze stellt eine entscheidende Herausforderung für die Entwicklung von effektiven Machine-Learning-Modellen im Bereich der *Computer Vision* dar. Diese Studie untersucht den Einfluss von synthetischen Daten auf die Leistung eines Mask R-CNN-Modells für die Detektion von Hautkrebs.

Um diesen Einfluss zu untersuchen, wurden synthetische Daten in das vorhandene Datenset integriert. Zwei verschiedene Ansätze wurden dabei verwendet: Zum einen wurden klassische Data-Augmentation-Techniken angewendet, bei denen Krebszellen als Vordergrund auf entsprechenden Hintergründen platziert wurden. Zum anderen wurden synthetische Bilder mithilfe von *Generative Adversarial Networks* (GANs) generiert.

Das Mask R-CNN-Modell wurde anschließend mit verschiedenen Datensätzen und Verhältnissen von synthetischen zu echten Daten trainiert. Die Leistung der trainierten Modelle wurde anhand eines separaten Testsets bewertet. Die Ergebnisse zeigen, dass die Ergänzung des Datensets mit synthetischen Daten die Leistung der Modelle verbessert. Besonders bemerkenswert ist, dass Modelle, die ausschließlich mit synthetischen Daten aus GANs erweitert wurden, die besten Leistungen erzielten.

Diese Erkenntnisse tragen dazu bei, das Verständnis dafür zu verbessern, wie synthetische Daten das Training von Machine-Learning-Modellen beeinflussen können, insbesondere im Bereich der medizinischen Bildgebung. Darüber hinaus legen sie nahe, dass die Verwendung von GANs eine vielversprechende Strategie für die Datenaugmentation in ähnlichen Anwendungen darstellen könnte.

Die Ergebnisse dieser Studie haben potenziell weitreichende Auswirkungen auf die Entwicklung von effektiven und robusten Machine-Learning-Modellen für die medizinische Diagnostik und anderer Einsatzgebiete.

Abstract

The availability of sufficient and diverse datasets poses a critical challenge for the development of effective machine learning models in the field of computer vision. This study investigates the impact of synthetic data on the performance of a Mask R-CNN model for skin cancer detection.

To examine this impact, synthetic data was integrated into the existing dataset. Two different approaches were employed: firstly, classical data augmentation techniques were applied, where cancer cells were placed as foreground on appropriate backgrounds. Secondly, synthetic images were generated using Generative Adversarial Networks (GANs).

Subsequently, the Mask R-CNN model was trained with various datasets and ratios of synthetic to real data. The performance of the trained models was evaluated using a separate test set. The results indicate that augmenting the dataset with synthetic data improves the performance of the models. Particularly noteworthy is that models exclusively augmented with synthetic data from GANs achieved the best performance.

These findings contribute to an improved understanding of how synthetic data can influence the training of machine learning models, especially in the field of medical imaging. Furthermore, they suggest that the use of GANs may be a promising strategy for data augmentation in similar applications.

The results of this study have potentially far-reaching implications for the development of effective and robust machine learning models for medical diagnostics and other domains.

Contents

1 Introduction **1**
- 1.1 Background and Motivation . . . 1
- 1.2 Problem statement and research objectives . . . 2
- 1.3 Research questions and hypotheses . . . 2
- 1.4 Overview on the structure of the thesis . . . 3

2 Foundations **4**
- 2.1 Introduction to Machine Learning and Image Processing . . . 4
- 2.2 Fundamentals of Artificial Neuronal Networks . . . 4
  - 2.2.1 Underlying mechanics of Artificial Neuronal Networks . . . 5
- 2.3 Data augmentation techniques for machine learning . . . 8
- 2.4 Fundamentals of Generative Adversarial Networks (GANs) . . . 9
  - 2.4.1 Underlying mechanics of conditional GANs . . . 10
- 2.5 Mask R-CNN . . . 12
  - 2.5.1 Underlying mechanics of CNNs . . . 13
  - 2.5.2 Underlying mechanics of Mask R-CNNs . . . 15

3 Related Work **17**
- 3.1 Review of relevant literature and studies . . . 17
- 3.2 Summary of findings and research gaps . . . 24

4 Methodology **25**
- 4.1 Description of the research approach . . . 25
- 4.2 Defining the Evaluation Criteria for Model Accuracy . . . 25
- 4.3 Mathematical Representation of Research Question . . . 26
- 4.4 Data Acquisition . . . 27
- 4.5 Explanation of applied data augmentation techniques and GANs . . . 27
- 4.6 Training of Mask R-CNN model . . . 28

5 Experiments **30**
- 5.1 Disclaimer . . . 30
- 5.2 Description of conducted experiments . . . 30
- 5.3 Development of Workflow . . . 31
- 5.4 Execution of tests and evaluation of models . . . 31

6 Results **32**
- 6.1 Presentation and discussion of results . . . 32
- 6.2 Interpretation of results in the context of research questions . . . 34

7 Conclusion **40**
- 7.1 Summary of Key Findings . . . 40
- 7.2 Implications of the Results for Practice and Future Research . . . 41

7.3 Future Work 41
7.3.1 Model Interpretability 41
7.3.2 Workflow adaptability 42
7.3.3 Optimizing Model Convergence Strategies 43
7.3.4 Unlocking Multi-Class Capabilities 43
7.3.5 Increase diversity of synthetic dataset 44
7.3.6 Alternatives to GANs 44

1 Introduction

1.1 Background and Motivation

The integration of Machine Learning (ML) models into the field of image analysis has brought about significant and positive changes, leading to substantial advancements in various sectors and industries. ML-powered vision systems have proven to be highly effective in enhancing the ability of autonomous vehicles to understand and interpret their surroundings. In complex scenarios, ML models decode intricate traffic situations with remarkable precision, improving overall safety and efficiency [5]. In industrial settings, machine learning models play a crucial role in quality control processes. By carefully identifying even the most subtle anomalies in manufactured components, these models ensure that products meet strict standards, thus maintaining high quality across production lines [44].

In the field of healthcare, the synergy of advanced imaging techniques and ML algorithms has empowered medical professionals. By detecting subtle deviations, these technologies facilitate early diagnoses and enable timely interventions, ultimately saving lives and improving patient outcomes [11]. Nevertheless, within these impressive advancements, a substantial challenge emerges - the requirement for extensive and diverse image datasets to train machine learning models effectively. The scarcity of suitable training instances has the potential to impede models from realizing their full potential, given the pivotal role of data in the learning process.

To address this challenge, a current strategy revolves around augmenting existing image repositories through deliberate image synthesis tailored specifically for training. This approach contains a spectrum of techniques, ranging from fundamental geometric transformations to Generative Adversarial Networks (GANs). Each method brings its own advantages and limitations, contributing to the complexity of the process.

Geometric transformations, such as rotation, scaling, and translation, offer computational efficiency and quick dataset expansion. However, while they provide rapid augmentation, they may not fully capture intricate real-world details. On the other hand, techniques like flipping and cropping diversify existing images and are widely used in deep learning. Despite their ability to enhance dataset diversity, they are constrained by the limitations of the original dataset, thus unable to generate entirely new perspectives.

On a more advanced level, GANs overcome those limitations by creating highly realistic synthetic images, so they excel in capturing complex patterns and generating novel content. Yet, they demand substantial computational resources and carefully tuning to prevent biases and maintain realism [33].

In handling this complex scenarios, the integration of synthetic images alongside authentic training data emerges as a pivotal solution. This integration not only addresses the challenge of dataset size but also resonates deeply with the inherent diversity of real-world scenarios. By empowering models with adaptability, this

fusion enables them to navigate diverse contexts with enhanced proficiency. Additionally, it reduces potential performance variations across different domains, ensuring consistent and reliable results in various applications. As technology continues to advance, finding innovative solutions to these challenges remains essential, ensuring the effective utilization of synthetic images in shaping the future of image analysis and Machine Learning.

1.2 Problem statement and research objectives

The use of synthetic data in machine learning pipelines to enhance model performance is a current practice that requires a careful exploration. This thesis delves into the difficult challenge of measuring the impact of synthetic data infusion on model accuracy, focusing on the context of skin cancer detection.

To start, a Mask R-CNN model's accuracy is precisely evaluated using a labeled skin cancer dataset from kaggle as the benchmark. A systematic augmentation approach is then undertaken, gradually introducing synthetic data.

Beginning with conventional data augmentation techniques, the study progressively transitions to synthetic data generated via GANs. This study provides a comprehensive examination of the intricate relationship between synthetic data and model accuracy, offering insights into the strategic amalgamation of diverse data augmentation techniques for practical implementation.

These findings offer valuable insights into the complex interplay between synthetic and real data, providing essential knowledge for making informed decisions when optimizing model efficacy, especially within the intricate domain of medical image analysis.

1.3 Research questions and hypotheses

Within this thesis project, an investigation into two fundamental research questions was made, to evaluate the impact of integrating synthetic data into the training process of Mask R-CNN models for skin cancer detection and clarify the nuances of data augmentation techniques and their role in model optimization:

1. *How does the integration of synthetic data, impact the accuracy and generalization capabilities of Mask R-CNN models in the context of skin cancer detection?*

It is hypothesized that the incorporation of synthetic data will lead to an improvement in the accuracy and generalization of Mask R-CNN models. By diversifying the training dataset, synthetic data is expected to enhance the model's ability to recognize subtle patterns and variations in skin cancer images, resulting in higher accuracy and improved generalization to unseen data. The introduction of synthetic data is anticipated to refine the model's understanding of complex skin cancer patterns, ultimately contributing to superior performance in both accuracy and generalization metrics.

2. *What is the comparative effectiveness of traditional data augmentation techniques versus synthetic data augmentation using GANs and how does the performance of Mask R-CNN models vary with the gradual infusion of synthetic data?*

We hypothesize that while traditional data augmentation techniques such as rotation, flipping, and scaling contribute to the model's performance, the gradual introduction of synthetic data will have a more substantial impact on enhancing accuracy. As the proportion of synthetic data in the training set increases, we expect a more pronounced improvement in performance. Synthetic data, by capturing diverse and intricate features, is anticipated to play a pivotal role in refining the model's understanding of complex skin cancer patterns. Consequently, the performance enhancement is expected to be more significant with the incorporation of synthetic data, demonstrating the importance of synthetic data augmentation in optimizing Mask R-CNN models for skin cancer detection.

1.4 Overview on the structure of the thesis

The thesis will follow this structure: Following the introduction, the foundational aspects of Machine Learning and Image Processing will be explored. A comprehensive examination of Mask R-CNN and cGANs will follow, utilizing the frameworks outlined in the respective papers for experimentation purposes. Following this, an introduction to related studies will be provided, encompassing various papers related to synthetic images, medical imaging, and Mask R-CNNs. A brief description of the methodology will outline the technical approach to the problem, followed by an exposition of the experiments conducted.

The results of the experiments will be presented and linked to the research question. Lastly, in the conclusion, the key findings will be summarized, and the research questions will be addressed.

2 Foundations

2.1 Introduction to Machine Learning and Image Processing

The blending of Machine Learning (ML) with image processing represents a pivotal point in today's technological landscape, offering transformation possibilities across diverse applications. This chapter lays the foundation for understanding the symbiotic relationship between these two domains, setting the stage for the subsequent exploration of their intricacies, methodologies, and real-world implications.

Machine Learning, a sub-field of artificial intelligence, equips systems with the capacity to learn patterns from data and make informed decisions without explicit programming. This autonomy grants computers the ability to refine their performance iteratively and tackle complex problems that elude conventional algorithms. ML's versatility has propelled it into various sectors, including finance, healthcare, and autonomous systems [42].

Image processing, on the other hand, centers on the manipulation and analysis of visual data, valuable insights from images. From medical imaging to satellite imagery, this field empowers computers to extract meaning and draw inferences from visual content. By leveraging computational techniques, image processing enhances the extraction of information, enabling applications such as object recognition, biometric identification and more [34].

The interplay between ML and image processing amplifies their individual capabilities, creating a dynamic alliance. ML brings data-driven reasoning to image analysis, while image processing augments the quality and relevance of data, enhancing the efficacy of ML models. This synergy manifests in diverse applications, from medical diagnoses [50] relying on image patterns to self-driving cars interpreting road scenes [7].

Nonetheless, this synergy introduces challenges. Processing large volumes of visual data demands sophisticated ML techniques to ensure efficiency. The complexities of image content, such as variations in lighting, orientation, and occlusions, require adaptable and robust ML models. Ethical considerations, encompassing privacy, fairness, and bias, emerge as critical aspects in this dynamic convergence.

Ethical considerations loom large, prompting the examination of bias, fairness, and privacy concerns [8]. The importance of ethical practices in designing ML-enhanced image processing systems that operate transparently and equitably should be mentioned as well.

In conclusion, the intersection of machine learning (ML) and image processing goes beyond their individual components, marking the onset of an era characterized by heightened intelligence [37].

2.2 Fundamentals of Artificial Neuronal Networks

Artificial neural networks (ANNs) were inspired by the impressive computational abilities of the human brain, which far surpass those of conventional digital comput-

ers. The brain's ability to perform complex, nonlinear, and parallel processing tasks, such as accurate predictions and pattern recognition, is facilitated by its network of neurons. Human vision, for instance, can process complex scenes and recognize patterns in a fraction of a second, a task that even powerful computers struggle to replicate. ANNs aim to mimic this efficiency and flexibility, offering promise in various fields requiring rapid and nuanced information processing [27].

The outstanding efficiency and speed of the human and animal brain function has long captivated researchers. Through a complex network of neurons, sensory signals from various bodily sensors are swiftly conveyed and processed, enabling rapid responses to both internal and external stimuli. Neurons, acting as autonomous units, transmit electrical signals along axons to synapse with dendrites of other neurons. With an estimated 100 billion neurons working in parallel, the brain's processing power lies in the distributed nature of neural activity and synaptic connections [17]. The amount of information processed and stored is influenced by the firing thresholds and the weighting assigned to each input by individual neurons. This intricate neural architecture underpins the brain's ability to perform intricate tasks efficiently and in near-instantaneous time frames [26].

ANNs are designed to emulate the functioning of the human brain, comprising hundreds or thousands of artificial neurons or processing units. They operate through computational learning algorithms that enable them to learn from experience rather than being explicitly programmed with rules. ANNs function as massively parallel computing systems, with interconnected neurons that learn and adapt based on environmental stimuli. Synaptic weights within the network capture and store the connections' strengths. The learning algorithm adjusts these weights sequentially and under supervision to achieve specific objectives. Neurons working collaboratively can learn intricate linear and nonlinear input-output relationships through sequential training methods. While ANNs were inspired by a different paradigm than statistical models, they share similar foundational elements. Some researchers assert that ANNs are essentially generalized nonlinear statistical models, even though with the complexity often hidden from users, making them more accessible to non-experts [28].

2.2.1 Underlying mechanics of Artificial Neuronal Networks

To explain the main elements used in an ANN, a general ANN model that contains this type of models is shown in figure (1). In an ANN, inputs ($x_1, ...x_p$) serve as the information fed into the network either from the surrounding environment or from interconnected neurons within the system. These inputs are then weighted by synaptic weights ($W = w_1, ...w_p$), which determine the influence of each input on the connected neurons. W represent in the case of a one-layer neural network the vector of the synaptic weights. In a multi-layer neural network W represents a matrix. Essentially, synaptic weights act as modulators, amplifying or attenuating the incoming information. b_j is known as the bias of a neuron that can act as a

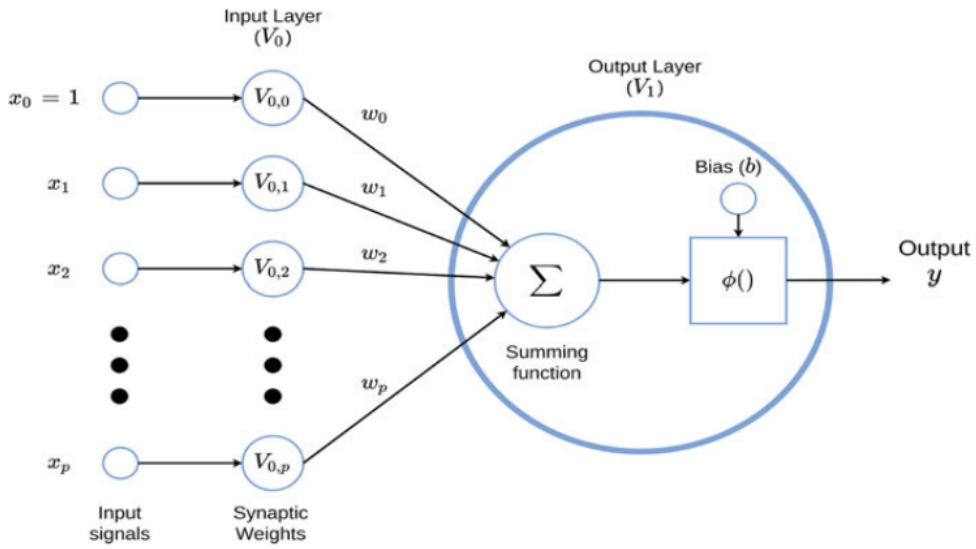

Figure 1: Artificial deep neural network [28]

threshold.

The net input v_j to a neuron is the aggregate result of the products of its inputs and their corresponding synaptic weights. It is computed by summing the products of input values and their respective synaptic weights that is mathematically described in 1.

$$v_j = \sum w_{ij} x_j \tag{1}$$

The activation function (g) of a neuron determines whether it should be activated based on the net input it receives. A common activation function is the rectified linear unit(ReLU) function. It introduces non-linearity into the network. The output of a neuron is determined by applying the activation function to its net input:

$$y_j = g(v_j) \tag{2}$$

This function introduces usually non-linearity into the network, enabling it to be applied in a large diversity of real problems.

Example 2.1. In Figure (2) a more complete picture of an ANN is shown. Since it has 2 hidden layers it can be described as a Deep Learning Model [29]. It is shown, that an ANN is characterized as a directed graph, where nodes represent neurons and edges represent connections between them [30]. In this structure, each neuron receives a weighted sum of the outputs from neurons connected to its incoming edges. The 2 layers (V_1,V_2) represent the hidden layers while V_0 is the input layer and V_3

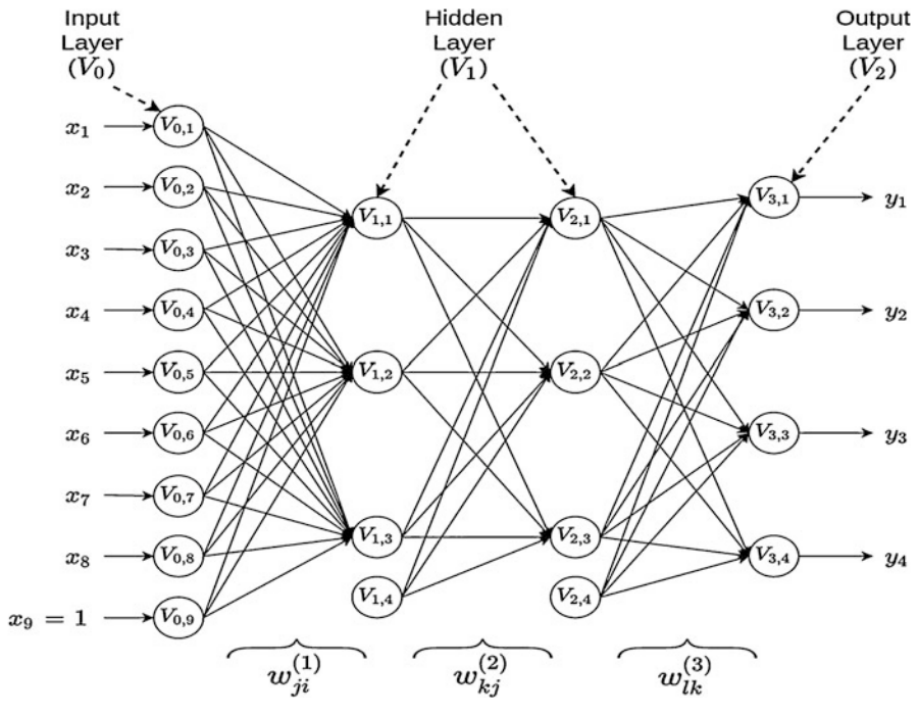

Figure 2: Artificial deep neural network [30]

denotes the output layer. The "depth" of the ANN can be described as three since V_0, which includes the input information, is excluded. The size of the network can be described as $|V| = 9+4+4+4 = 21$ in each layer a '+1' is added to the observed units to represent the node of the bias. The width of the network is $max|V_t| = 9$. The analytical form of the model shown for output o with d inputs, M_1 hidden neurons in hidden layer 1, M_2 hidden units in hidden layer 2, and O output neurons is given by the following:

$$V_{1j} = g_1 \sum_{i=1}^{D} w_{ji} x_i \quad for \quad j = 1, ..., M_1 \tag{3}$$

$$V_{2k} = g_2 \sum_{j=1}^{M_1} w_{kj} V_1 j \quad for \quad k = 1, ..., M_2 \tag{4}$$

$$V_{1j} = g_3 \sum_{k=1}^{M_1} w_{lk} V_2 k \quad for \quad l = 1, ..., O \tag{5}$$

The output of each neuron in the first hidden layer is generated by (3), while (4) computes the output of each neuron in the second hidden layer. Finally, (5) determines the output of each response variable of interest. The learning process involves

the utilization of weights ($w_{ji}^{(1)}$, $w_{kj}^{(2)}$, $w_{lk}^{(3)}$), which are organized in the following vector:

$$w = (w_{11}^{(1)}, w_{12}^{(1)}, ..., w_{1d}^{(1)}, w_{21}^{(2)}, w_{22}^{(2)}, ..., w_{2M_1}^{(2)}, w_{31}^{(3)}, w_{32}^{(3)}, ..., w_{3M_2}^{(3)}) \tag{6}$$

g_1 and g_2 denote the activation functions in the hidden layers, while g_3 represents the activation function of the output layer. The model is structured into interconnected layers: the input layer, hidden layers, and output layer. Each layer performs nonlinear transformations through artificial neurons, and connections between these layers are established using weights. In cases where only one output variable is present, the model is referred to as a univariate DL model.

If only one hidden layer exists, the DL model reduces to a conventional artificial neural network model. However, the inclusion of more than one hidden layer allows for the better capture of complex interactions, nonlinearities, and nonadditive effects [31].

2.3 Data augmentation techniques for machine learning

Data augmentation serves as a fundamental tool within the field of computer vision, enriching the performance and resilience of machine learning models through the introduction of diverse training data [24]. This introduction explores the significance, methodologies, and practical applications of data augmentation techniques, laying the groundwork for a deeper comprehension of their role in various computer vision tasks.

Computer vision, which aims to enable machines to understand visual information, has made significant progress, particularly with the rise of deep learning [2]. However, despite these advancements, the limited availability of extensive labeled datasets can hinder model effectiveness. Data augmentation addresses this issue by artificially enlarging datasets through diverse transformations, thereby improving the model's capacity to generalize [24].

The spectrum of data augmentation strategies encompasses a variety of techniques that modify existing images while preserving their underlying semantics. Common methods include rotation, translation, scaling, flipping, and adjustments to brightness and contrast. These controlled variations enable the model to capture the inherent diversity present in real-world images.

The implications of data augmentation are manifold. In tasks such as object recognition, augmenting data with diverse viewpoints and orientations aids in model robustness [2]. For medical imaging, transformations simulate anatomical variations, facilitating better performance in different scenarios. Augmentation is also pivotal in scenarios with class imbalances, addressing the challenge of limited samples for underrepresented classes.

The effectiveness of data augmentation relies on its careful implementation. Factors such as the level of augmentation, choice of transformations, and the blend of

augmented and original data play a crucial role in shaping the model's learning capabilities. Achieving this balance necessitates both practical experimentation and a deep understanding of the domain [1].

In the upcoming sections, a comprehensive exploration of various data augmentation techniques will be undertaken. This includes the augmentation with traditional augmentation techniques as well as the generation of synthetic Images using Generative Adversarial Networks.

Furthermore, domain-specific augmentations will be considered, with techniques tailored to different application areas such as medical imaging data analysis. The significance of augmentation in addressing challenges such as variations in lighting, occlusions, and background complexities will be underscored.

While ethical considerations may not be directly tied to augmentation, responsible usage ensures that models generalize effectively across diverse scenarios, contributing to unbiased outcomes by minimizing inadvertent biases introduced by limited training data.

In conclusion, data augmentation emerges as a pragmatic approach for increasing the capabilities of machine learning models in computer vision. This section establishes a foundation for the exploration of an array of augmentation methods, preparing the ground for a deeper dive into methodologies, challenges, and practical implementations that will be used during the experiments.

2.4 Fundamentals of Generative Adversarial Networks (GANs)

In recent years, the landscape of image-based applications has undergone significant transformations, attributed in large part to the advancements in Generative Adversarial Networks (GANs).

Machine learning algorithms, such as artificial neural networks, have found extensive applications in image recognition, drug discovery, self-driving cars, and more. This summary introduces GANs, a type of ML algorithm comprising a discriminator and a generator engaged in a competitive learning process. The discriminator aims to distinguish between real and generated samples, while the generator tries to produce realistic images. GANs are widely utilized in diverse applications, including human face generation, image inpainting, face aging, image super-resolution, anime character generation, style transfer, and medical image analysis [21]. The summary outlines GAN fundamentals, contemporary variants, application examples, and discusses challenges in GAN implementation, emphasizing the dynamic competitive process between the generator and discriminator that culminates in an ideal Nash equilibrium point, making GANs a potent tool for image generation [21].

Besides GANs facilitate image-to-image translation, enabling the transformation of input images into outputs with distinct attributes. Notable methods like Pix2Pix and conditional GANs(cGans) excel in translating semantic maps, labels, and even altering day to night. Style transfer is another application where images retain their

structure while adopting the style from another source, resulting in effects like converting photographs into paintings or altering architectural textures [9]. A variety of use cases for image-to-image translation are shown in figure (3).

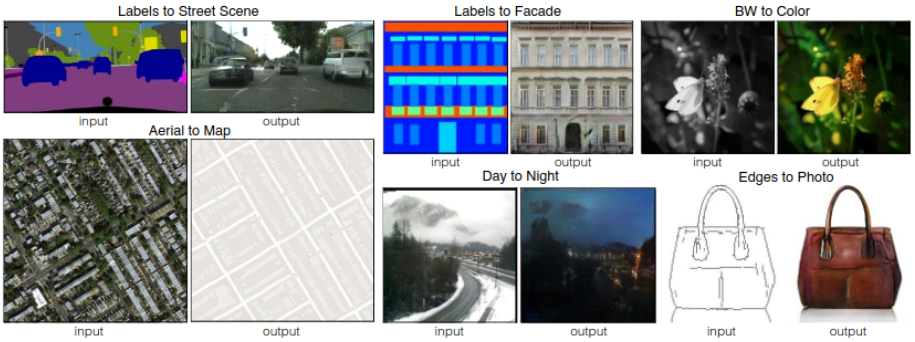

Figure 3: Examples Image to Image translation [19]

evIn summary, the emergence of GANs marks a significant paradigm shift in machine learning, empowering the creation of data that stretches beyond conventional boundaries of imagination and reality. This section serves as the starting point for a detailed examination of GANs, facilitating a deeper understanding of their architectural complexities, training intricacies, and transformative influence on tasks such as image synthesis, manipulation, and augmentation.

2.4.1 Underlying mechanics of conditional GANs

In this section, we delve into the underlying mechanics of GANs and conditional GANs(cGANs). The generator-discriminator dynamic and the adversarial training process that drives data synthesis will be shown and the mathematical background explained.

GANs are generative models that learn a mapping from a random noise vector z, element from a space Z, to an output image y, element from a space Y. z is a randomly generated vector sampled from a probability distribution. It serves as a source of randomness for the generator network. Formally, this can be represented as a function $G : Z \rightarrow Y$, where $G(z) = y$.

In contrast in cGANs, the generation of the image y is additionally influenced by a condition x. This condition x could be, for example, an image that instructs the Generator to produce an image of a specific class. Formally, we consider the Generator of a cGAN as a function $G : X \times Z \rightarrow Y$, where X is the set of possible conditions and Z is the space of noise vectors. So, the Generator takes a pair (x, z) as input and generates the corresponding image y.

The generator G is trained to produce outputs that cannot be distinguished from "real" images by an adversarially trained discriminator, D, which is a binary classifier function adjusted to do as well as possible at detecting the generator's "fakes".

Unlike an unconditional GAN, both the generator and discriminator observe the input x. The training procedure is diagrammed in figure (4).

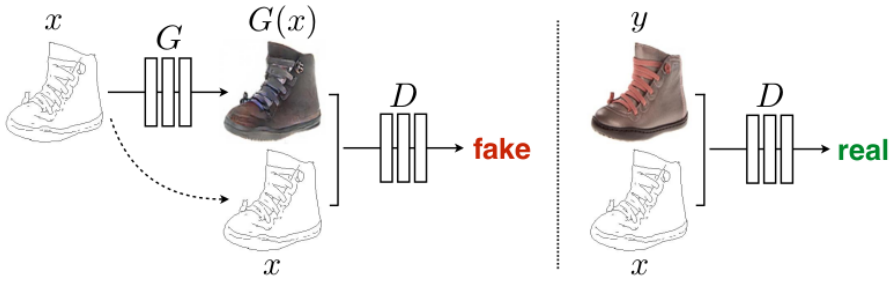

Figure 4: Training a conditional GAN to map edges to photos [19]

The objective of a conditional GAN can be expressed as:

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \tag{7}$$

where G tries to minimize this objective against an adversarial D that tries to maximize it. $\mathbb{E}_{x,y}[\log D(x, y)]$ represents the adversarial loss, that can be described as the expected value of the logarithm of the discriminator's output when given a real pair of (x, y). $\mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]$ represents the generator loss, that can be described as the expected value of the logarithm of 1 minus the discriminator's output when given a fake pair $(x, G(x, z))$. In these expressions, $\mathbb{E}_{x,z}$ and $\mathbb{E}_{x,z}$ denotes the expectations taken over all possible values of the variables x, y and x, z. Previous approaches have found it beneficial to mix the GAN objective with a more traditional loss, such as L2 distance [32]. The L2 distance, also known as the Euclidean distance, measures the pixel-wise squared differences between two images [10]. The discriminator's job remains unchanged, but the generator is tasked to not only fool the discriminator but also to be near the ground truth output in an L2 sense. It was explored that instead of using the L2 distance using the L1 distance is beneficial. The L1 distance, also known as the Manhatten distance measures the absolute pixel-wise differences between two images. The use of L1 distance can reduce the blurring in the generated images and is defined in (8) [19].

$$L_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \tag{8}$$

The final objective for a cGAN tested in the "Image-to-Image Translation"-Paper is:

$$G^* = arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G) \tag{9}$$

In summary, the objective is to find the optimal generator G^* that minimizes the combination of the conditional GAN loss and the L1 regularization term. The discriminator D is simultaneously trained to maximize the performance of the GAN. The hyperparameter λ controls the trade-off between the adversarial loss and the L1 regularization. This kind of objective is often used in image-to-image translation tasks, where the generator is trained to produce realistic outputs while maintaining a certain level of similarity to the target [19].

2.5 Mask R-CNN

The vision community has made rapid progress in object detection and semantic segmentation, driven by frameworks like Fast/Faster R-CNN and Fully Convolutional Network (FCN). The goal described in the Mask R-CNN paper [13] is to create a similarly enabling framework for instance segmentation.

Instance segmentation is challenging as it requires detecting all objects in an image and precisely segmenting each instance. It combines elements from object detection and semantic segmentation. Mask R-CNN extends Faster R-CNN by adding a mask prediction branch to the existing classification and bounding box regression branches. The mask branch is a small Fully Convolutional Network applied to each Region of Interest (RoI), predicting segmentation masks in a pixel-to-pixel manner. To address misalignment issues in Faster R-CNN caused by coarse spatial quantization during feature extraction, the authors propose RoIAlign, a quantization-free layer that preserves exact spatial locations. This change significantly improves mask accuracy [13].

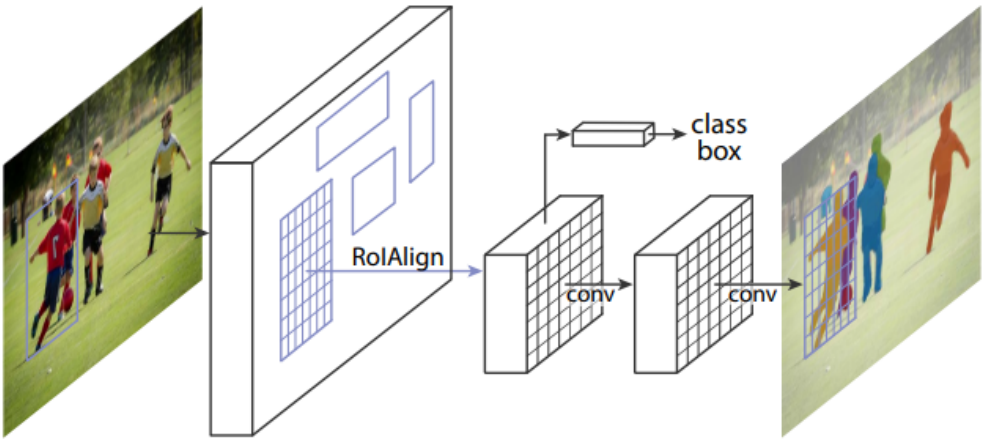

Figure 5: The Mask R-CNN framework for instance segmentation[13]

Unlike some other methods, Mask R-CNN decouples mask and class prediction [14]. It predicts a binary mask for each class independently and relies on the network's RoI classification branch to predict the category. Mask R-CNN outperforms previous state-of-the-art single-model results on the COCO instance segmentation task. The framework's flexibility and accuracy are highlighted, and the authors believe it will benefit and ease future research on instance segmentation [15]. The generality of the framework is demonstrated by applying it to the task of human pose estimation on the COCO keypoint dataset. It surpasses the winner of the 2016 COCO keypoint competition while running at 5 frames per second [14]. In summary, Mask R-CNN is introduced as a simple, flexible, and fast framework that achieves state-of-the-art results in instance segmentation, with particular emphasis on overcoming spatial quantization issues and decoupling mask and class prediction. The framework is showcased for its speed, accuracy, and versatility in handling different computer vision tasks [16].

2.5.1 Underlying mechanics of CNNs

CNNs are similar to traditional ANNs in that they consist of neurons that self-optimize through learning. Each neuron receives inputs and performs operations, such as scalar products followed by non-linear functions, similar to traditional ANNs. The entire network expresses a single perceptive score function, with the last layer containing loss functions associated with classes. Techniques developed for traditional ANNs apply to CNNs as well [25].

The main difference between CNNs and traditional ANNs is that CNNs are primarily used for pattern recognition within images. CNN architectures encode image-specific features, making them suitable for image-focused tasks and reducing the parameters required to set up the model.

One limitation of traditional ANNs is their struggle with the computational complexity of image data. For example, the MNIST dataset, with its small image dimensionality, is manageable for most ANNs. However, with larger, colored images, the number of weights in the network increases substantially, leading to computational challenges [3].

Overfitting is a significant concern in machine learning, including with ANNs. It occurs when a network cannot effectively learn due to various reasons. It is crucial to reduce overfitting's effects to improve the model's generalization performance . Reducing the complexity of ANNs helps mitigate overfitting by reducing the number of parameters required to train the network, thereby improving predictive performance [4].

CNNs consist of three main types of layers: convolutional layers, pooling layers, and fully-connected layers [25]. A simple CNN architecture for classification is shown in Figure (6). These layers are stacked to form the architecture of a CNN:

1. **Input Layer:** This layer holds the pixel values of the input image, similar to other types of Artificial Neural Networks (ANNs).

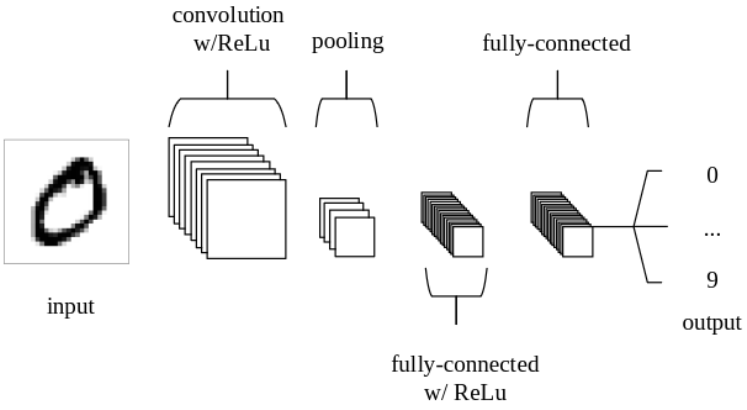

Figure 6: Example: CNN architecture [25]

2. **Convolutional Layer:** Neurons in this layer are connected to local regions of the input image. Each neuron calculates the scalar product between its weights and the region connected to the input volume. The Rectified Linear Unit (ReLU) activation function is commonly applied to the output of this layer.

3. **Pooling Layer:** This layer performs downsampling along the spatial dimensions of the input, reducing the number of parameters in the activation.

4. **Fully-Connected Layers:** These layers, similar to traditional ANNs, aim to produce class scores from the activations obtained from previous layers.

The ReLU activation function is a commonly used non-linear activation function in neural networks. It operates by setting all negative values in the input to zero, while leaving positive values unchanged. Mathematically, the ReLU function is defined as:

$$\text{ReLU}(x) = \max(0, x) \tag{10}$$

Essentially, the ReLU function introduces non-linearity by allowing only positive values to pass through unchanged, while effectively "turning off" negative values. This simple thresholding operation helps in addressing the vanishing gradient problem and accelerates the training of deep neural networks[6].

The convolutional layer, shown in Figure (7) is a key component of CNNs, responsible for learning and detecting features within input data. It operates by convolving learnable kernels across the spatial dimensions of the input, generating 2D activation maps. These kernels, which are small in spatial dimensionality but span the depth of the input, calculate scalar products as they slide across the input. Through

this process, the network learns to identify specific features at different spatial positions, leading to the generation of activations.

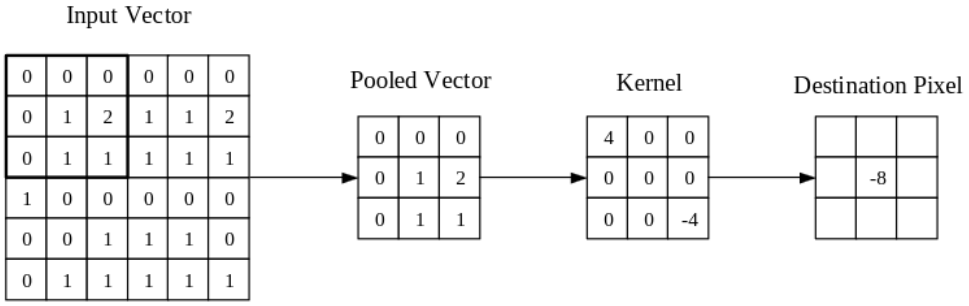

Figure 7: Visual representation of the convolutional layer [25]

2.5.2 Underlying mechanics of Mask R-CNNs

Mask R-CNN typically uses a backbone CNN architecture, such as ResNet or a similar feature extractor, to extract hierarchical features from the input image. Those features are used for region proposal and mask prediction.

Like Faster R-CNN, Mask R-CNN includes a Region Proposal Network (RPN) that proposes candidate bounding boxes likely to contain objects. The RPN is responsible for suggesting RoIs based on the hierarchical features obtained from the backbone network [39].

One key modification introduced by Mask R-CNN is RoIAlign. In traditional approaches like RoIPool, spatial quantization is applied during feature extraction, leading to misalignment between the input and output. RoIAlign is a quantization-free layer that accurately preserves spatial locations during feature extraction. This is crucial for precise pixel-wise segmentation [12].

Mask R-CNN inherits the classification and bounding box regression branches from Faster R-CNN. The classification branch assigns class probabilities to each RoI, and the bounding box regression branch refines the predicted bounding box coordinates.

The key addition in Mask R-CNN is the introduction of the mask branch. For each RoI, a small Fully Convolutional Network (FCN) is applied to predict segmentation masks. This branch produces a binary mask for each class independently in a pixel-to-pixel manner.

Mask R-CNN decouples the mask prediction from the class prediction. It predicts a binary mask for each class independently without competition among classes. The RoI classification branch is responsible for predicting the category [15].

- Classification Loss: Measures the difference between predicted class probabilities and ground truth class labels.

- Bounding Box Regression Loss: Measures the difference between predicted bounding box coordinates and ground truth coordinates.
- Mask Segmentation Loss: Compares pixel-wise binary mask predictions to ground truth masks using binary cross-entropy.

The loss of the Mask R-CNN can be defined as:

$$L = L_{Class} + L_{BBox} + L_{Mask}$$

Mask R-CNN is trained in a multi-task manner, optimizing the combination of classification, bounding box regression, and mask segmentation losses. The back-propagation of gradients allows the network to learn to simultaneously perform these tasks. During inference, the trained model can be used to detect and segment objects in new images. Thresholding is applied to obtain binary masks, and post-processing steps may be employed for refinement.

3 Related Work

3.1 Review of relevant literature and studies

Through a comprehensive review of existing papers published in several journals, an understanding of the current research progress in the field of ML-driven solutions in Computer Vision tasks has been achieved. Additionally, several papers concerning Data Augmentation methods have been examined. In the following section, the most important papers concerning those topics will be presented, and research gaps and opportunities will be identified.

A reviewed paper [38] addresses the challenge of accurate and timely interpretation of Chest X-rays (CXRs) due to limited medical resources, proposing a computer-aided diagnosis (CAD) system based on machine learning as a solution. However, acquiring a sufficiently large, balanced, and annotated dataset of CXRs for training such a system is challenging. The study conducts a comparative analysis on learning from imbalanced and limited CXRs to detect pneumonia, focusing on two main questions:

1. What is the effectiveness of data sampling methods in improving the performance of learning models?

2. What are the quantifiable differences between learning models that use different sampling techniques?

Two categories of data sampling techniques are explored: undersampling the majority class and oversampling/augmentation of the minority class. The study evaluates Support Vector Machine and deep convolutional neural network models, demonstrating that both exhibit improved performance when appropriate data sampling strategies are employed, based on experimentation with a publicly available CXR dataset.

In the data undersampling pipeline, the authors balance the class distribution by randomly sampling data from the majority class. The chosen handcrafted feature for disease detection in Chest X-rays (CXRs) is the histogram of oriented gradients (HoG), and a traditional machine learning model, Support Vector Machine (SVM), is employed for analysis. In contrast, the data oversampling strategy involves expanding the minority class using basic image transformations such as reflection and rotation. Additionally, data augmentation techniques utilizing GANs, specifically Deep Convolutional GAN (DCGAN), are employed due to its stable training capabilities. The architecture of CNN classifier is accordingly designed by the authors to achieve a balance between performance and training speed, reducing the risk of overfitting posed by very deep networks. Inspired by Zeiler and Fergus's net (ZFNet), a CNN architecture asssociated to ZFNet is developed to address this concern effectively. The ZFNet is a CNN architecture developed by Matthew D. Zeiler and Rob Fergus [48]. It gained prominence for its performance in the ImageNet Large-Scale Visual Recognition Challenge in 2013 [41].

The authors are evaluating the performance of their classifier using a range of metrics, including accuracy, specificity, sensitivity (recall), precision, F1-score, G-mean, area under the ROC curve (AUC), and area under the precision-recall curve (AUPRC). These metrics allow them to assess various aspects of the classifier's performance, such as its ability to correctly classify instances and its balance between precision and recall. They also highlight the importance of Precision-Recall curves, particularly in assessing classifier performance in imbalanced datasets.

Results indicated that without data sampling, the SVM achieved the highest recall but the lowest specificity due to heavy imbalance. However, implementing data sampling generally improved test accuracy, specificity, precision, and F1 score for the SVM model, with data undersampling showing the best improvement. Data oversampling methods yielded high recall but slight improvements in precision and specificity.

Comparing SVM and CNN performance, CNN classifiers showed better overall performance, with higher AUC values, indicating better feature representations. After data sampling, CNN performance further improved, with specificity scores increasing notably. Augmentation with GAN showed the best performance overall, with a high AUC. The overall Results are shown in Table 1 and Table 2.

Metrics	Data without augmentation	Data under sampling	with affine transformation	with GAN augmentation
Accuracy	0.724	0.817	0.777	0.750
Recall	0.987	0.872	0.933	0.946
Specificity	0.286	0.726	0.517	0.423
Precision	0.697	0.842	0.763	0.732
F1 score	0.817	0.856	0.840	0.826
AUC	0.898	0.897	0.896	0.891

Table 1: Performance of SVM with different data sampling methods

Metrics	Data without augmentation	Data under sampling	with affine transformation	with GAN augmentation
Accuracy	0.853	0.877	0.872	0.899
Recall	0.946	0.862	0.854	0.897
Specificity	0.697	0.902	0.902	0.902
Precision	0.839	0.936	0.935	0.938
F1 score	0.889	0.897	0.893	0.917
AUC	0.911	0.940	0.938	0.954

Table 2: Performance of CNN with different data sampling methods

For future work, the authors' intention is to investigate the proposed methods for

other datasets, where the challenge of limited and imbalanced data may be even more severe. This would likely reveal an even more indispensable need for data balancing techniques to achieve adequate performance. In particular, the gap between strategies without data sampling and those with data augmentation may be even more significant, i.e., the results shown in Tables 1 and 2 would show even more improvements when introducing data balancing, since the initial performance would be poor because of sensitivity to the effects of small data size and imbalances.

Since data augmentation within the field of medical imaging is examinated in multiple papers, a advanced approach for Data augmentation for CXR Dataset is described in another paper from the same authors[2]. Similar to the previous paper the authors mention the impressive progress for Deep Neural Network-based methods in the field of medical imaging tasks and the limited availability of large and well distributed Dataset. The authors tested their augmentation methods on a U-Net. A U-Net is a convolutional neural network architecture that is widely used for image segmentation tasks. It was introduced by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015 through their paper titled "U-Net: Convolutional Networks for Biomedical Image Segmentation."

The U-Net architecture is designed to address the challenge of pixel-wise classification, particularly in biomedical image segmentation tasks where precise delineation of structures like cells, organs, or tumors is crucial. It consists of a contracting path, which captures context via convolutional and pooling layers, followed by an expansive path, which enables precise localization through upsampling and convolutional layers [40]. The key features of U-Net include skip connections between the contracting and expanding paths as shown in Figure (8). This architecture has proven to be highly effective in various segmentation tasks to biomedical imaging.

As Data augmentation methods on the CXR image dataset the authors presented a data augmentation technique that "uses different combinations of contrast, brightness and gaussian filters, to imulate CXR images with extreme opacities and low contrast" [2]. To evaluate the effectiveness of this augmentation technique outperforms standard data augmentation techniques such as rotating, flipping and zooming. For the evaluation Dice coefficient(DC) and Intersection over Uninon(IoU) were used and defined as followed with G for the ground truth mask and P for the segmented lung mask. Both values are matrices where each element corresponds to a pixel in the image. Each element has a binary values. Where a 1 indicates that the pixel does belong to the object of interest and a 0 does not belong to the object of interest.

$$DC(G,P) = \frac{2 \cdot |PG|}{|P| + |G|} \tag{11}$$

$$IoU(G,P) = \frac{|PG|}{|P| + |G| - |PG|} \tag{12}$$

The training was run on a HPC network with 2 x Intel Xenon 6130 CPU, 192GB Ram, and 4 x Nvidia V100 GPUs. For the three public dataset used the results are

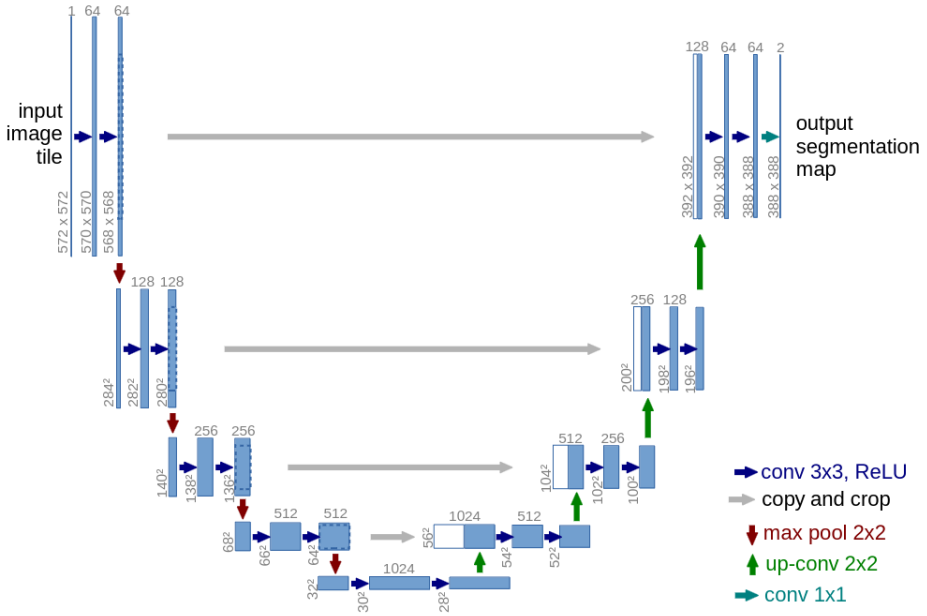

Figure 8: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations [40].

shown in Table (3).

Dataset	Augment Type	DC (mean $\pm$ std. %)	IoU (mean $\pm$ std. %)
Dataset 1	Std aug	0.9579 ± 2.92	0.9206 ± 5.08
	Prop aug	0.9879 ± 0.54	0.9761 ± 1.04
	Std + Prop aug	0.9613 ± 2.30	0.9264 ± 4.08
Dataset 2	Std aug	0.8744 ± 9.23	0.7875 ± 13.13
	Prop aug	0.9494 ± 2.32	0.9046 ± 4.05
	Std + Prop aug	0.8935 ± 7.33	0.8149 ± 11.19
Dataset 3	Std aug	0.9294 ± 5.40	0.8724 ± 8.78
	Prop aug	0.9495 ± 3.19	0.9055 ± 5.51
	Std + Prop aug	0.9326 ± 5.85	0.8787 ± 9.24

Table 3: Performance Metrics for Different Augmentation Types

The proposed augmentation(Prop aug) achieved on all 3 Datasets better results in DC and IoU compared to the standard augmentation(Std aug) and a mix of standard

and proposed augmentation methods.

Another relevant paper [49] introduces a new evaluation method for the detection and classification of wind turbine blade defects. The increasing demand for wind power has led to a rise in the inspection and repair of wind turbine blades (WTBs). Defect detection systems can be employed to ensure the ongoing performance and maintenance of WTBs. The paper investigates the performance of deep learning algorithms — specifically YOLOv3, YOLOv4, and Mask R-CNN — in detecting and classifying defects in wind turbine blades (WTB). YOLO (You Only Look Once) was developed by Redmon and Farhadi from the University of Washington, used for object detection based-segmentation [35].

Besides traditional evaluation measures based on precision, recall and the F1-socre the authors introduces novel performance evaluation measures tailored for defect detection tasks, including Prediction Box Accuracy (14), Recognition Rate (15), and False Label Rate (16) all depended on the Bounding Box Accuarcy (14), which is the mean of the mean of the width accuary and the height accuary of the bounding box. A bounding box refers to a rectangular box drawn around an object or area of interest in an image that contains a defect(9).

$$\text{Bounding Box Accuracy(BBA)} = \frac{WidthAcc + HeightAcc}{2} \tag{13}$$

$$PBA = \frac{1}{N}\sum_{i=1}^{n} \text{BBA}_i \tag{14}$$

$$RR = \frac{1}{N}\sum_{i=1}^{n} 1, \text{ if BBA}_i > 0 \tag{15}$$

$$FLR = \frac{1}{N}\sum_{i=1}^{n} 1, \text{ if } (\text{BBA}_i > 0) \wedge (\text{Predicted Type}_i \neq \text{Labelled Type}_i) \tag{16}$$

Experiments conducted on a dataset provided by an industrial partner, comprising images from WTB inspections, revealed that Mask R-CNN consistently outperformed other algorithms, especially with transformation-based augmentations like rotation and flipping. Specifically, using the best dataset, Mask R-CNN achieved a mean Weighted Average (mWA) value of 86.74%, surpassing YOLOv3 (70.08%) and YOLOv4 (78.28%).

Furthermore, the authors introduce the Image Enhanced Mask R-CNN (IE Mask R-CNN) pipeline, which integrates optimal combinations of image enhancement and augmentation techniques, along with a tuned Mask R-CNN model, to address challenges in WTB defect detection. In conclusion, the paper presents an investigation into the performance of deep learning algorithms for WTB defect detection and classification. Using a dataset provided by Railston & Co. Ltd., the study explores the impact of various image augmentation and enhancement techniques on algorithm performance.

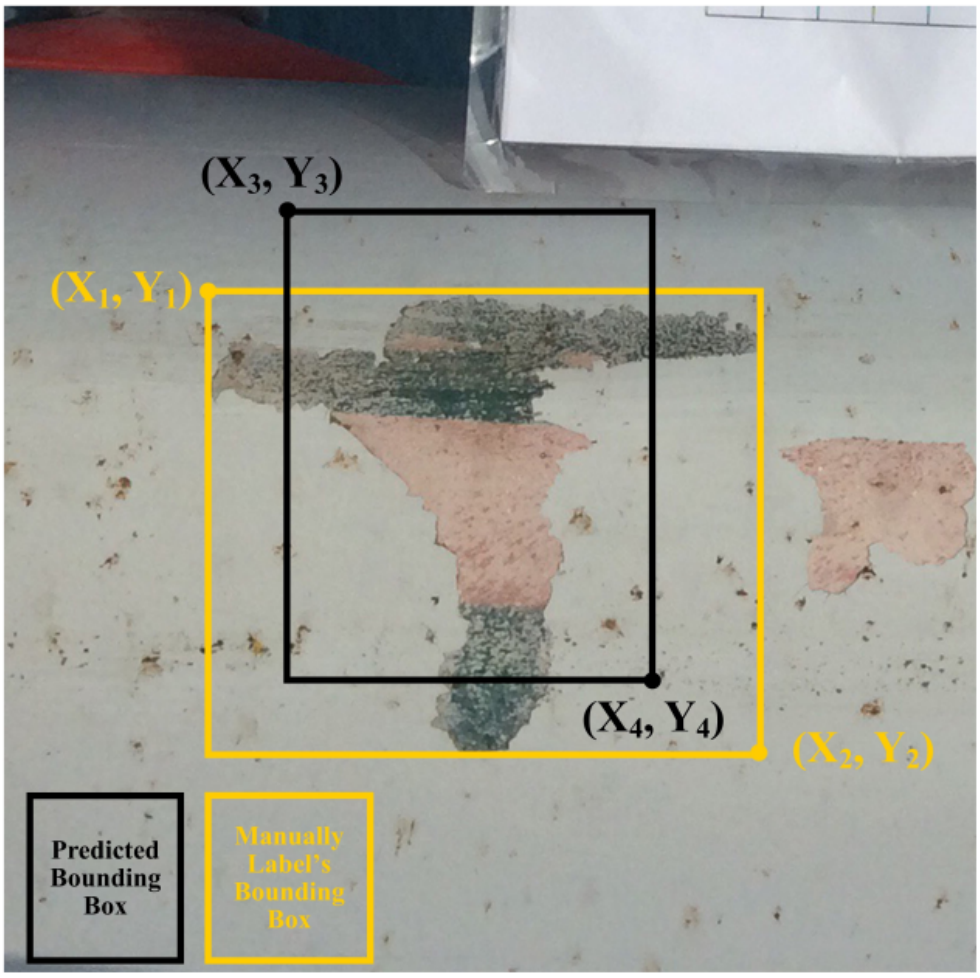

Figure 9: Example: predicted and real bounding box[49]

New evaluation measures, including Prediction Box Accuracy (PBA), Recognition Rate (RR), and False Label Rate (FLR), offer a comprehensive assessment of defect detection performance. Results indicate that Mask R-CNN outperforms other algorithms, particularly with transformation-based augmentations. Looking ahead, the paper suggests exploring additional image enhancement techniques and optimizing CNN parameters for different detection scenarios. The importance of condition monitoring and fault diagnosis for WTBs is underscored, with implications for the broader field of defect detection systems based on machine learning and deep learning methods. In summary, the proposed Image Enhanced Mask R-CNN pipeline holds promise for WTB defect detection and offers potential applications in various defect detection scenarios beyond WTBs.

The paper "Mixing Real and Synthetic Data to Enhance Neural Network Training - A Review of Current Approaches" [43] addresses challenges in computer vision caused by the need for large annotated datasets. The authors are giving insides into the recent advancements in computer vision that enabled machines to achieve human-level accuracy in tasks like object detection and classification. It is mentioned that some domains face challenges in collecting extensive data, especially when rare events or privacy concerns are involved and the primary limitation to the performance of CNNs is the scarcity of training data rather than the network architecture itself. While large-scale datasets exist for certain applications, annotating them requires significant effort. Because of those difficulties the authors suggest the usage of synthetic data.

Synthetic data offers benefits such as cost-effectiveness and ease of annotation, but it must accurately reflect the feature distribution of real-world data to avoid performance gaps caused by domain shift. This issue is particularly relevant in domains like urban and traffic scenes, where data collection can be challenging. The paper focuses on reviewing CNN training scenarios that enhance network performance without relying on additional real-world data, particularly in urban and traffic scene contexts.

Besides the generation of synthetic data the paper points out the possibility to use pretrained networks that are mostly trained on large datasets such as ImageNet, Pascal VOC or COCO. The scale of these datasets enables neural networks to acquire abstract representations across numerous object categories and incorporate generalized feature representations. When customizing a neural network for a specific task, the dataset size specific to that task is often smaller compared to the dataset used for initial training. Therefore, it's important to assess which existing network architecture best suits the desired accuracy and available computational resources before proceeding with customization.

Increasing reliance on synthetic data during training increase the challenge of domain shift, although there's evidence suggesting that utilizing more synthetic data can reduce its effects, leading to well-performing networks on real data. In semantic segmentation tasks, synthetic data might suffice for training background classes covering large image areas. However, for foreground classes representing individual objects, synthetic data often lacks realistic textures, necessitating a training approach focused on object detection rather than semantic segmentation. Further research in this area is encouraged to gain deeper insights.

In 2020, when the paper was released, the authors expected the usage of Generative Adversarial Networks (GANs) as a fast growing research area, with expectations of more datasets emerging. These datasets could feature automatically generated, photo-realistic urban scenes where relevant instances like pedestrians or cars are inserted in various poses and random positions within the images. Looking at the performance and the distribution of today's available image generators the authors expectation for the increased usage of GANs was met [23].

3.2 Summary of findings and research gaps

In the field of medical imaging, leveraging Generative Adversarial Networks (GANs) for chest X-ray (CXR) datasets enhances convolutional neural network (CNN) performance in lung disease classification. Recent studies addresses challenges by exploring image quality effects, proposing a multi-scale CNN architecture, and analyzing different learning models using GAN-augmented images.

Shifting to wind turbine blade inspections, researcher from industrial fields compared deep learning algorithms, highlighting Mask R-CNN's superior accuracy. Future research can explore challenges associated with extensive labeled data and computational resources in this context.

If it comes to image-to-image translation,researchers introduced Conditional Adversarial Networks (cGANs) as a versatile solution, showcasing effectiveness in various tasks. Future research should delve into potential drawbacks or challenges in real-world applications. For computer vision challenges with large annotated datasets, "Mixing Real and Synthetic Data to Enhance Neural Network Training" explores methods to create powerful neural networks with less reliance on real-world data. Future research opportunities include a detailed exploration of domain shift's impact on model performance and strategies to mitigate this challenge.

These studies represent significant advancements in their domains and although it offers future research opportunities include a comprehensive examination of how various methods of creating synthetic data impact model performance. This research gap warrants exploration to enhance the understanding and optimization of synthetic data generation techniques.

4 Methodology

4.1 Description of the research approach

In the assessment of the Mask R-CNN model's effectiveness in skin cancer segmentation, a systematic research methodology is employed. As a basis a skin cancer dataset sourced from Kaggle [22] is utilized, with a primary focus on segmentation tasks. The dataset is independently labeled using *Supervisely*, an online tool for labeling, with the accompanying CSV file set aside, prioritizing segmentation aspects over classification. Supervisely provides a range of annotation tools for labeling objects in images or video frames, including bounding boxes, polygons, points, and other annotation types. The "smart label tools" offered by Supervisely significantly reduced the time required for labeling tasks [45].

In the next step, to enrich the labeled dataset, two distinct augmentation strategies are implemented. The first involves traditional techniques like scaling and rotation to introduce variations. The second explores the creation of synthetic images through the use of cGAN. The usage of two different augmentation strategies aims to inject diversity and complexity into the dataset for a more comprehensive analysis of the Mask R-CNN model's performance.

Following data preparation, multiple iterations of the Mask R-CNN model are trained with varying ratios of synthetic images, starting with the initial model that is exclusively trained on the original labeled images as a baseline for performance. Subsequent models progressively incorporate augmented and synthetic images in varying proportions. This staged augmentation process systematically evaluates the model's adaptability to augmented and synthetic data, uncovering potential improvements in segmentation accuracy.

The performance of each CNN with respect to segmentation accuracy is evaluated based on a custom metric, as explained in the next section. This comparative analysis tries to show trends, identify potential trade-offs, and determine optimal scenarios for leveraging augmented and synthetic data in the context of skin cancer segmentation.

The research approach, characterized by precise data preparation, staged model training, and stringent evaluation, provides a structured framework for investigating the nuanced interplay between traditional and synthetic data augmentation techniques in enhancing the Mask R-CNN model's performance for skin cancer segmentation tasks.

4.2 Defining the Evaluation Criteria for Model Accuracy

As a basis for an evaluation metric, two intermediate metrics the F1-Score as well as the Intersection over Union (IoU) were calculated. The calculation is done on an unseen testset and through the various models.

IoU measures the overlap between the predicted mask and the ground truth mask. It is calculated as the ratio of the area of intersection between the predicted

and ground truth masks to the area of their union. Higher IoU values indicate better alignment between the predicted and ground truth masks.

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{17}$$

The F1 score involves comparing individual pixels between the predicted masks and the ground truth masks. Each pixel in the predicted mask is compared with the corresponding pixel in the ground truth mask to determine whether it represents a true positive, false positive, true negative, or false negative prediction. These comparisons are used to compute precision, recall, and ultimately the F1 score, which provides a measure of the model's performance in identifying defects accurately while minimizing false identifications and missed defects. The F1 score ranges from 0 to 1, with higher values indicating better model performance. It serves as a comprehensive evaluation metric for defect recognition systems, capturing both the ability to correctly identify defects and the ability to avoid false identifications, based on the pixels provided as input.

$$F1 = \frac{\text{2 x Precision x Recall}}{\text{Precision + Recall}} \tag{18}$$

with Precision and Recall defined as following:

$$Precision = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{19}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{20}$$

4.3 Mathematical Representation of Research Question

The hypothesis posits that augmenting the dataset with synthetic data leads to increased values in both the F1-score and the IoU in the Mask R-CNN. This assumption is grounded in previous research indicating that enhancing the diversity and quantity of training data can enhance the model's ability to generalize to unseen examples. The equal importance to both IoU and F1-scores in the evaluation is assigned. Therefore, the evaluation score ES is defined as shown in Equation (21).

$$ES = \frac{\text{IoU + F1}}{2} \tag{21}$$

Assuming that both IoU and the F1 score depend on the ratio of synthetic and original data added to the training set, we introduce a ratio function f. This results in Equation (22):

$$ES(f) = \frac{\text{IoU(f) + F1(f)}}{2} \tag{22}$$

Where f represents the ratio of original to synthetic images, defined as:

$$f = \frac{D_{syn}}{D_{orig}} \tag{23}$$

Here, D_{orig} represents the number of original images, and D_{syn} represents the number of synthetic images..

4.4 Data Acquisition

Obtaining the dataset involved uploading 1200 images from the Kaggle Dataset to Supervisely for labeling tasks. Following the labeling process, Supervisely provided the data in the form of a Common Objects in Context (COCO) Dataset. The COCO dataset primarily focuses on object recognition and detection, featuring images with multiple objects in complex scenes, with each object annotated with a bounding box. In addition to bounding box annotations, the COCO dataset includes pixel-wise segmentation masks for object instances [20].

4.5 Explanation of applied data augmentation techniques and GANs

In the generation of synthetic images using traditional data augmentation methods, masks from the COCO dataset will be utilized to extract defects from the original images. This extraction process will result in skin cancer cells depicted against a transparent background, which will serve as foregrounds. Additionally, a selection of images depicting healthy skin without cancer cells will be curated to serve as backgrounds. The synthetic skin cancer cells will undergo augmentation through random flipping, resizing, and subtle alterations to color attributes. These augmentation techniques aim to simulate the natural variations observed in real skin cancer cells, thereby enriching the dataset with diverse visual characteristics. After accumulating a sufficient number of augmented skin cancer cells, the next step involves their random merging with the background images. This merging process ensures that each synthetic image presents a unique composition of foreground and background elements. Concurrently, as these new composite images are created, corresponding masks outlining the boundaries of the cancerous regions will be generated. The composite images and the corresponding masks are the foundation of the Mask R-CNN training. Figure 10 and figure 11 are showing results of both generation processes.

In order to create synthetic images via a cGAN architecture, the proposed pix2pix framework [19] was used. The original images and corresponding masks were utilized to train the pix2pix model. After training the synthetic Masks that were already created with the traditional methods were fed them into the generator. This should result in more realistic looking images.

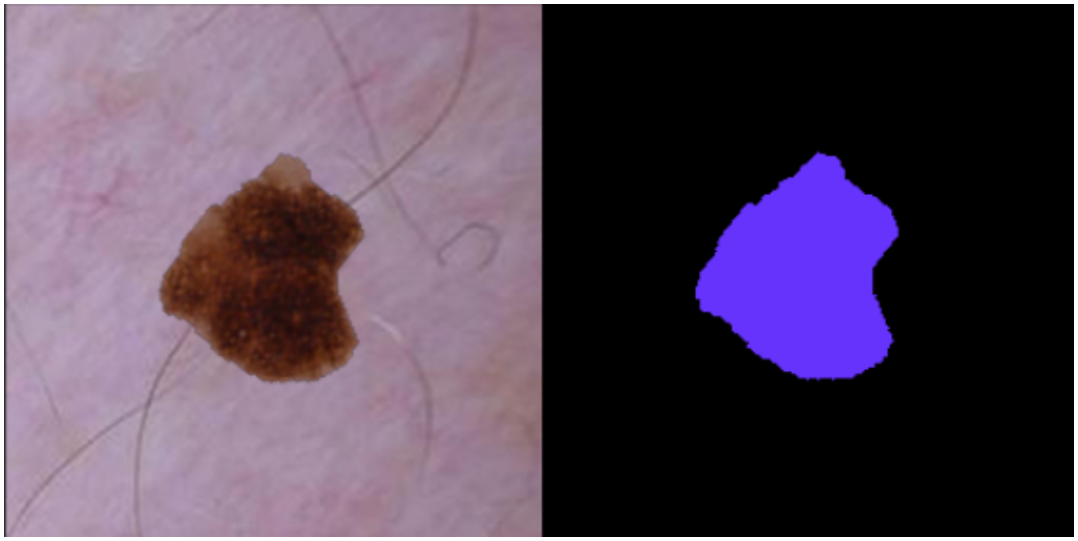

Figure 10: Example: Composite Image and Mask

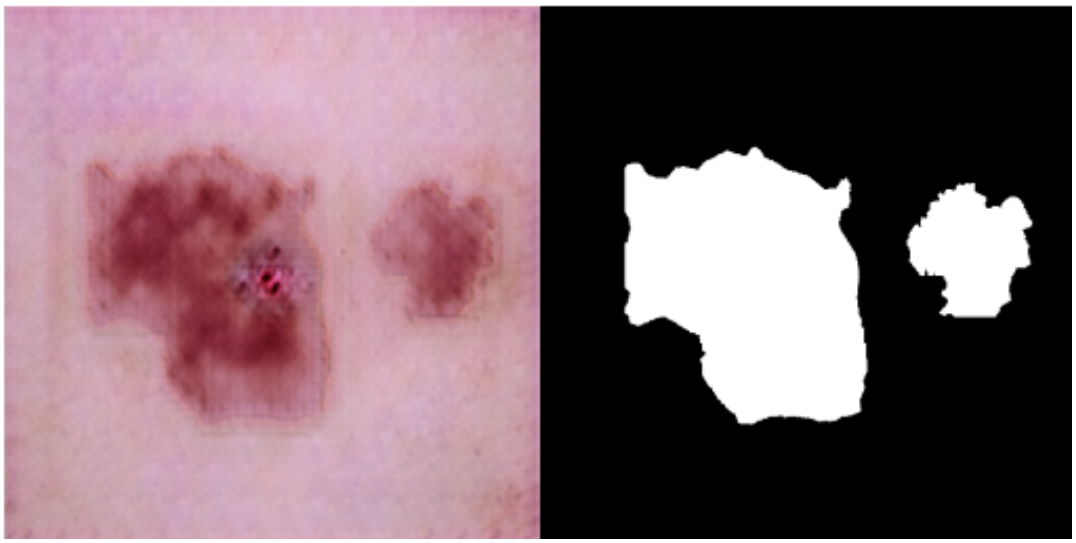

Figure 11: Example: Image and Mask from cGAN

4.6 Training of Mask R-CNN model

The first step in training the Mask R-CNN model using PyTorch involves data preparation. This includes curating a dataset comprising images of skin lesions annotated with bounding boxes and segmentation masks. The dataset is divided into training and validation sets to facilitate model training and evaluation. The Mask R-CNN model is initialized with a pre-trained backbone network, typically ResNet50, which is pre-trained on large-scale image datasets such as COCO (Common Objects

in Context). The backbone network provides feature representations that are leveraged by the Mask R-CNN architecture for object detection and instance segmentation. GPU acceleration was used for enhanced training and inference performance. PyTorch, a popular deep learning framework, is utilized for model development and training due to its flexibility, ease of use, and extensive community support [36].

Data loaders are employed to efficiently load batches of images and corresponding annotations during training. Data augmentation techniques, including random flips, rotations, and color jittering, are applied to augment the training dataset, thereby enhancing the model's generalization ability and robustness to variations in input data. The loss function used for training the Mask R-CNN model comprises multiple components, including classification loss, localization loss, and segmentation loss. These components collectively optimize the model parameters to accurately predict object classes, localize object bounding boxes, and segment object masks.The training loop iterates over multiple epochs, with each epoch consisting of multiple iterations over batches of training data. During each iteration, forward and backward passes are performed to compute the loss and update the model parameters using the Stochastic Gradient Descent (SGD) optimization algorithms.

5 Experiments

5.1 Disclaimer

The skin cancer dataset utilized in this study is employed exclusively for benchmarking and research purposes. It is important to note that the approaches developed and discussed in this thesis are intended for scientific exploration and computational analysis only. They are not designed, nor should they be interpreted, to provide medical advice or diagnostics. The methodologies presented here do not replace professional medical judgment, diagnosis or treatment.

5.2 Description of conducted experiments

The baseline training involves training the Mask R-CNN model using the original training dataset of 1000 images without synthesis or augmentation. This serves as the initial training step to establish a benchmark performance for the model without any additional data manipulation or techniques. After completing the baseline training, the performance of the model is evaluated on an unseen dataset of 200 images using the IoU and the F1 score. The baseline training serves as a reference point for subsequent experiments, allowing for comparison with models trained using synthetic data or data augmentation techniques. The following table shows the combination of real and synthetic images used in the conducted experiments:

Nr	Dataset Size	Real Images	Synthetic Images	GAN Images
1	1.000	1.000	0	0
2	3.000	1.000	2.000	0
3	3.000	1.000	0	2.000
4	3.000	1.000	1.000	1.000
5	15.000	1.000	14.000	0
6	15.000	1.000	0	14.000
7	15.000	1.000	7.000	7.000

Table 4: Summary of Conducted Experiments

These experiments aim to address the two main research questions. Experiments 2-4 will assess the impact on a model when the dataset is augmented with synthetic images that are twice the size of the original data. Meanwhile, experiments 5-7 will investigate whether significantly enlarging the dataset yields any benefits. Additionally, the experiments will compare the performance of two synthetic image generation methods: brute force placement of cancer cells on skin versus the use of a cGAN for synthetic image generation.

5.3 Development of Workflow

To streamline the training process, a comprehensive workflow designed to simplify data input and model training is developed. The primary objective of this workflow is to input the original images and the annotation from supervisly in the COCO annotation format. A preprocessing script uses the annotation file to create to each image an corresponding mask. Besides this every labled defect is extracted as a foreground. The original images and their masks are resized to 256x256 pixel to increase the batch size due to limited memory on the GPU.

After the preprocessing a baseline dataset is created. It holds the extracted foregrounds, some backgrounds and the original resized images and their corresponding masks. In the next step, the total dataset size and the desired number of each type of synthetic images is specified. The workflow employs a brute-force method to generate randomized images and masks using the provided foregrounds and backgrounds. Subsequently, the original images and their corresponding masks are utilized to train the cGAN using the Pix2Pix framework previously mentioned.

Upon completion of the cGAN training, the randomized masks generated by the previous launched brute-force method are the input into the Pix2Pix transition to produce synthetic images. A shuffle script is then applied to create the final image and mask dataset comprising brute-force, synthetic, and cGAN-generated images. The resulting dataset is fed into the Mask R-CNN training for 20 epochs. The computational workload is offloaded from the CPU to an NVIDIA RTX 4090 GPU using CUDA, which enables a batch size of 8 and significantly reduces training time. Specifically, training 3000 images takes 12 hours, while training with 15,000 images extends to 50 hours. The workflow itself is published in a GitHub Repository [47].

For visual reference, an illustration of the workflow is provided in figure 12.

5.4 Execution of tests and evaluation of models

Throughout an extensive training process for approximately 200 hours across seven different models, the GAN underwent periodic resets and complete retraining alongside each iteration of the Mask R-CNN model. This method guaranteed a fair comparison among the different models, enhancing their overall robustness and accuracy in the final assessments. Moreover, to maintain data integrity and consistency, a secondary script was written to oversee the proper composition of the training dataset.

Following the training process, each model underwent evaluation over 20 epochs, with the loss recorded and plotted for analysis. Beyond assessing the Mask R-CNN loss, the models were subjected to testing using an independent script. This script coordinated the assessment of the models' performance using 200 new images from the skin cancer dataset. By conducting thorough analysis, it calculated the IoU and F1 score, along with their respective standard variations. These metrics serve as indicators of the models' accuracy and ability to generalize, and they were recorded at the end of the testing script.

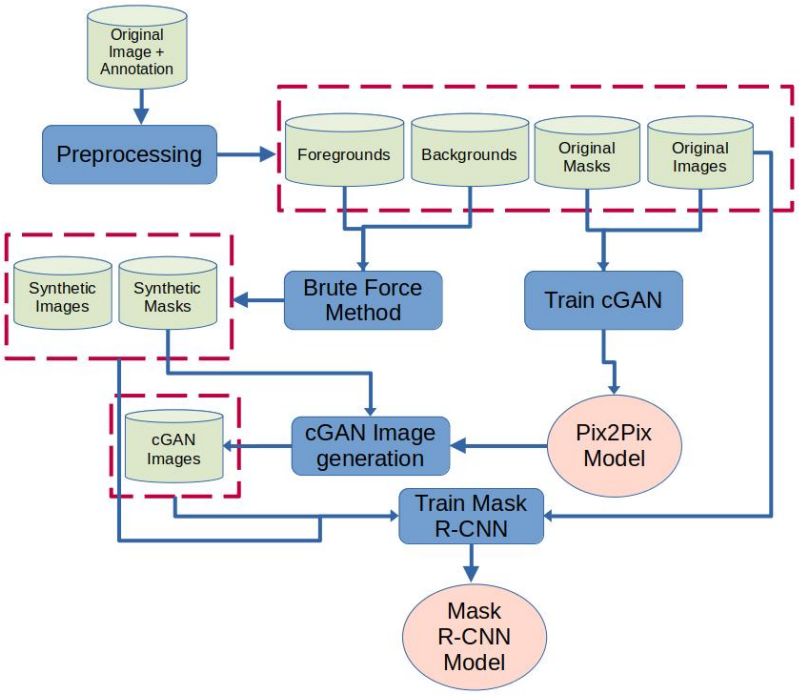

Figure 12: automated workflow for image generation and Mask R-CNN training

6 Results

6.1 Presentation and discussion of results

During training the Mask R-CNN showed a very rapid stabilization and a minimal Loss. After 10 epochs stabilized around 0.05. Pre-trained models, especially those like ResNet50 trained on large-scale image datasets like ImageNet, have already learned to extract useful features from images. As a result, when fine-tuning such a pre-trained model on a specific dataset, it can converge faster and achieve lower loss compared to training from scratch. The development of the loss shown in the 15.000 GAN Dataset is shown in figure 13.

The original dataset, comprising 1.000 images, served as the baseline for our experiments. The Mask R-CNN model trained on this dataset achieved an IoU of 0.2286 and an F1 score of 0.2941. Augmenting the dataset to 3000 images, both through traditional methods and GAN augmentation, led to notable performance improvements. The GAN-augmented trial saw significant enhancement, with IoU increasing to 0.6329 and F1 to 0.7399. This highlights the effectiveness of GAN-generated data in improving model accuracy.

While the trial employing traditional augmentation methods also led to perfor-

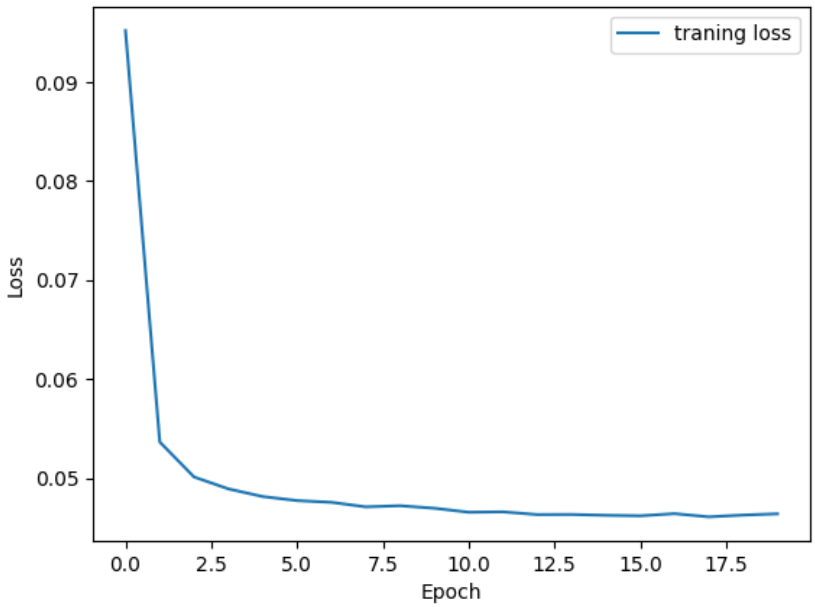

Figure 13: Development of Loss over Epoch during Mask R-CNN training

mance gains, with IoU increasing to 0.3546 and F1 to 0.4157, the improvement was less pronounced compared to GAN augmentation. Scaling up the dataset to 15,000 images, both with traditional augmentation methods and GAN augmentation, further improved model performance. The GAN-augmented trial showed remarkable enhancement, with IoU increasing to 0.7552 and F1 to 0.8458, showcasing the effectiveness of GAN-generated data, particularly on larger datasets. While traditional augmentation methods also improved performance on the 15,000-image dataset, with IoU increasing to 0.3246 and F1 to 0.3970, the gains were less substantial compared to GAN augmentation. The model trained on the hybrid datasets with a mix of GAN-generated data and traditional augmented data performed better then the ones trained only on traditional augmented datasets, but worse then the ones trained exclusively on the GAN-generated data.

All experiment results including the evaluation score and the evaluation matrix is shown in table 5. The bar chart provided in Figure 14 allows an easier comparison of the models.

The evaluation matrix indicates that training Mask R-CNN with 15.000 GAN-generated images yields the highest success rate in detecting cancer cells in the unseen validation dataset. Subsequently, predictions of various models are visualized to demonstrate how mask predictions vary with different dataset sizes and augmentation methods. Figure 15 illustrates the predictions of three different Mask R-CNN models on an image from the validation dataset. While the model trained on the original dataset fails to detect any cancer cells in the predicted mask, both models

Table 5: Performance Metrics

Experiment	Mean IoU	Std IoU	Mean F1	Std F1	ES
Original Dataset (1.000)	0.2286	0.2865	0.2941	0.3376	0.2614
GAN (3.000)	0.6329	0.2548	0.7399	0.2286	0.6864
Traditional (3.000)	0.3546	0.3666	0.4157	0.4017	0.3852
Hybrid (3.000)	0.6050	0.2707	0.7115	0.2547	0.6582
GAN (15.000)	0.7552	0.1826	0.8458	0.1429	0.8005
Traditional (15.000)	0.3246	0.3391	0.3970	0.3680	0.3608
Hybrid (15.000)	0.7187	0.2272	0.8097	0.2051	0.7642

trained on 15.000 images successfully identify cancer cells. The GAN-augmented model demonstrates superior precision in outlining the contours of the marks. Figure 16 presents a comparison between models trained with 3.000 and 15.000 images, highlighting the advantages of the larger dataset. While both models can detect cancer cells, the 3.000-image model incorrectly interprets hairs as cancer cells. While the model trained on the 15.000 GAN-generated images typically produces the best results, there are exceptions in the validation dataset. Figure 17 illustrates how both the model trained on the original images and the one trained on 3.000 augmented images identify cancer cells. However, the model trained on the 15.000 images not only detects cancer cells but also misidentifies the black background in the bottom right corner as a defect.

This can be explained by overviewing the images produced by the generator. Those look more realistic than the ones with the standard augmentation methods. But a lack of variety can be identified. It seems that the cGAN created only one type of skincolor and a similar type of cancer cells. This led to the issue that all images generated by the cGAN have a similar appearance, which is shown in Figure 18.

6.2 Interpretation of results in the context of research questions

In this thesis project, two pivotal research questions concerning the effectiveness of integrating synthetic data and employing various augmentation techniques on the performance of Mask R-CNN models in the realm of skin cancer detection were addressed.

The first question revolves around understanding how the incorporation of synthetic data influences the capabilities of Mask R-CNN models. Our results demonstrate a significant impact on model performance when synthetic images are introduced. Specifically, the the incorporation of synthetic images led to an improved Mask R-CNN in every case.

The second question delves into the impact of different augmentation techniques on performance. Two methods were investigated: firstly, overlaying cancer cells onto clear skin images using a brute-force approach, and secondly, training a GAN

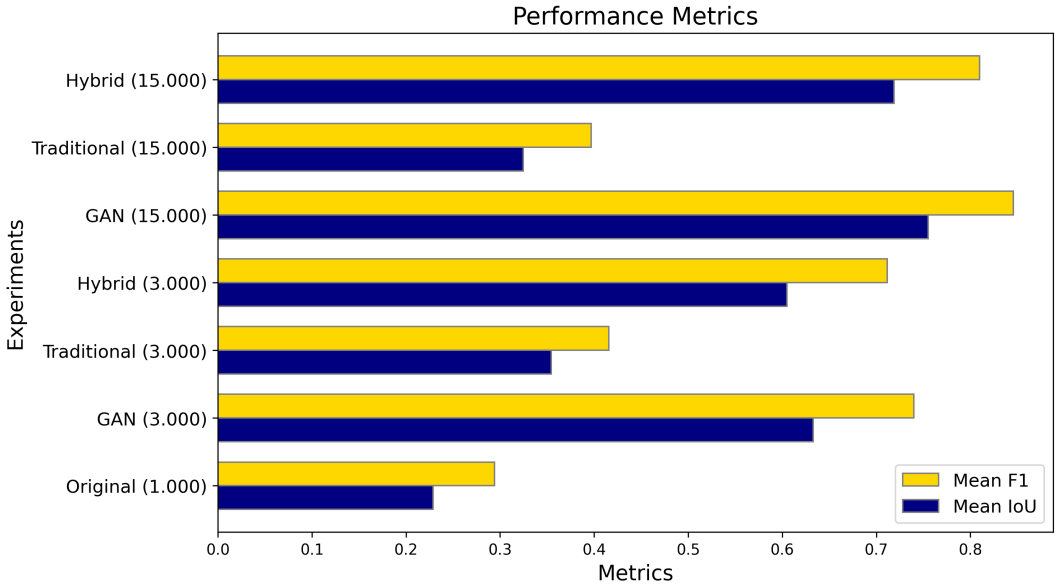

Figure 14: Performance Chart of all Experiments

on the original dataset and utilizing its generator to augment the dataset for Mask R-CNN training. Our findings reveal that the method utilizing images generated by the GAN yielded significantly higher IoU and F1 scores. However, while these scores were superior for certain images in the validation dataset, other models achieved better overall results.

Chapter 4.3 introduces the Evaluation Score (ES), defined as the mean of the IoU and the F1-score, and postulates a dependency on the ratio of the IoU and the F1-score. Figure 19 showcases the fitting of a logistic function (24) to the data points of the model trained solely on GAN-generated images. The logistic function was fitted to the results by using the $curvefit$ function of the $SciPy$ library. "SciPy provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics and many other classes of problems."[46]

$$f(x) = \frac{0.8}{1 + e^{-1.259(x-0.575)}} \tag{24}$$

Logistic functions are frequently utilized to model growth or diffusion processes constrained by various factors. Although the limited number of models trained, may not provide sufficient data points to conclusively support whether this logistic function accurately represents the development of the Evaluation Score, it can be inferred that even with a larger volume of synthetic images, the Evaluation Score will plateau.

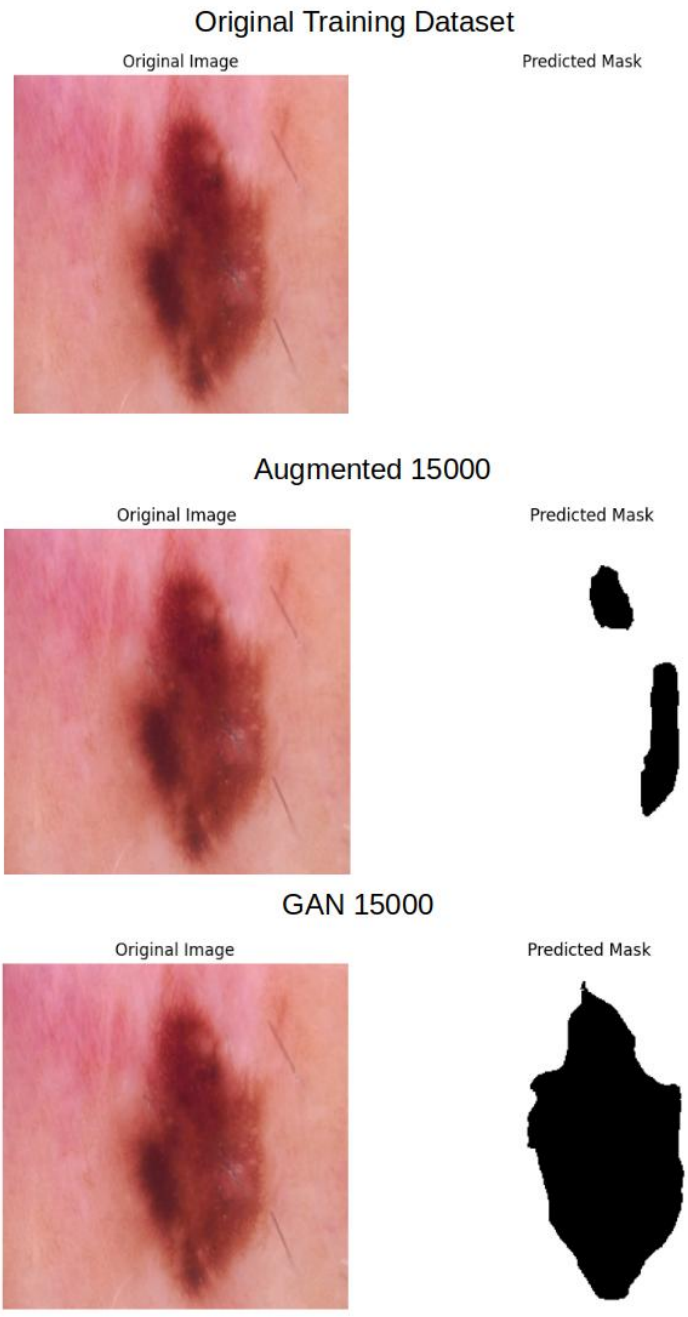

Figure 15: Prediction of various models on image from validation dataset

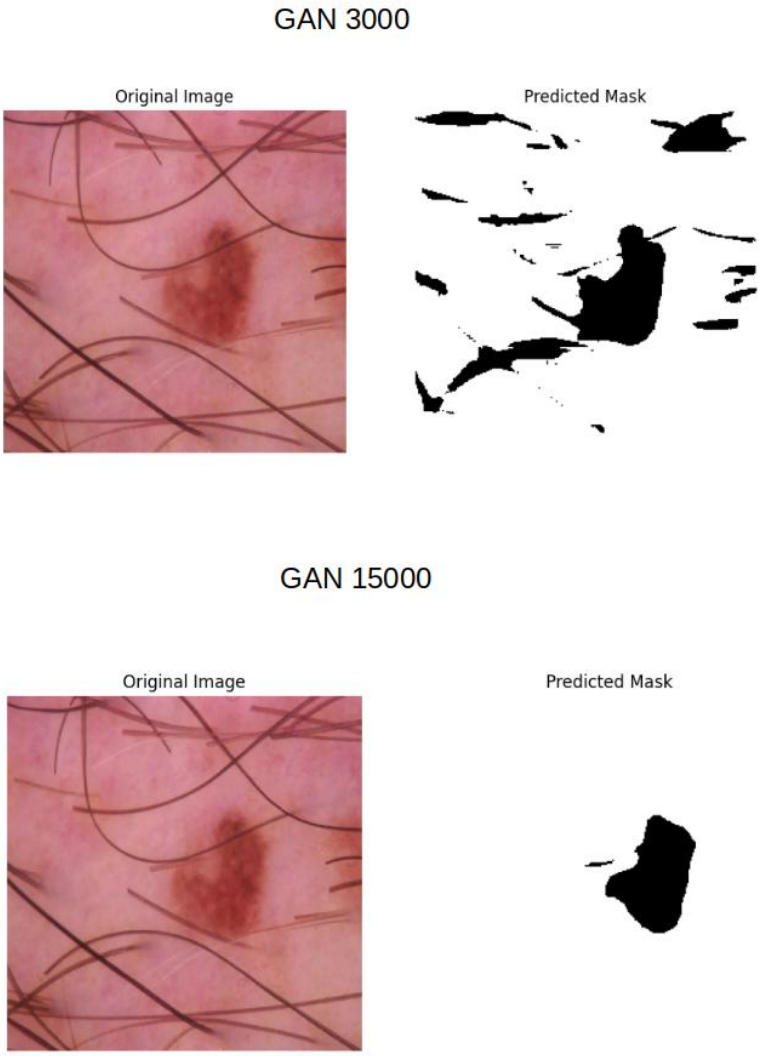

Figure 16: Comparison of models trained on 3.000 and 15.000 images

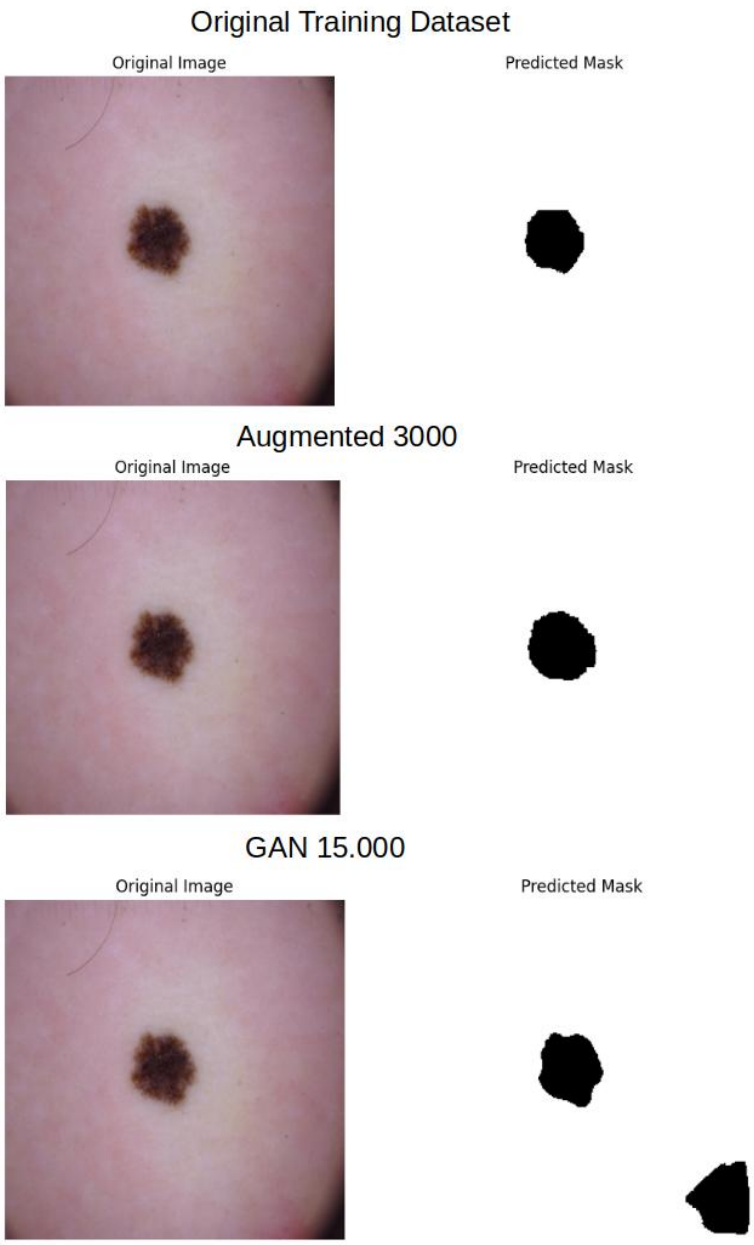

Figure 17: Displaying detecting issues due to an unbalanced training dataset

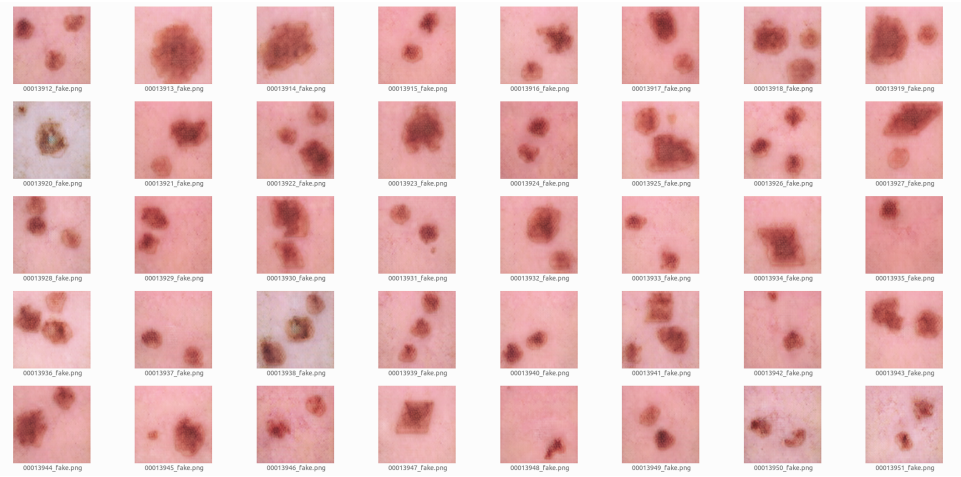

Figure 18: Displaying a extract from the cGAN generated images, that indicates a lack of variety

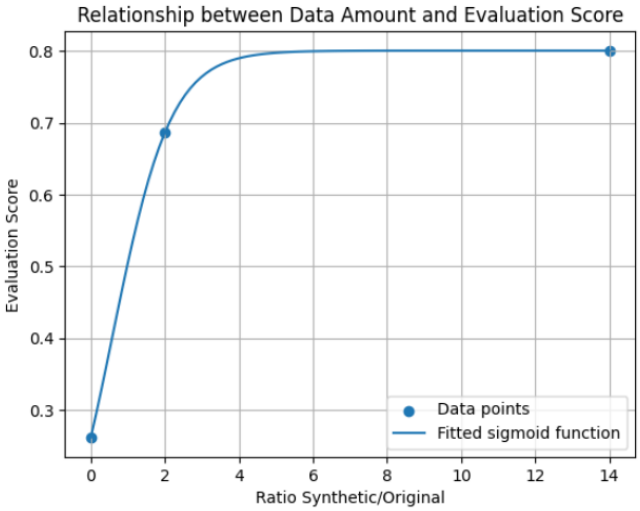

Figure 19: Relationship between Synthetic/Original Ration and Evaluation Score

7 Conclusion

7.1 Summary of Key Findings

Data augmentation plays a crucial role in enhancing the accuracy of Mask R-CNN models. However, the effectiveness of augmented data can vary depending on the method employed and the specific application. During the investigation using the skin cancer dataset, we discovered that generating images with a cGAN produced superior results compared to traditional augmentation techniques. Nonetheless, it's important to acknowledge that the generated data exhibited a lack of diversity, indicating that further diversification of the dataset could potentially enhance model performance.

Another key finding underscores the critical importance of professionally labeled and high-quality baseline datasets for training and evaluation purposes. While labeling tools like Supervisely offer efficiencies in the annotation process, they may inadvertently introduce biases into the model. To mitigate this risk, it is imperative to actively involve domain experts with deep knowledge and understanding of the data domain, particularly in medical imaging where precision is paramount for accurate diagnoses and treatment decisions. Engaging domain experts, such as experienced radiologists or dermatologist, can ensure that the baseline datasets are labeled with the highest level of accuracy and consistency. Their expertise in recognizing subtle patterns and nuances in medical images can help identify and correct potential biases or errors introduced by automated labeling tools, resulting in more reliable and trustworthy models.

Furthermore, collaborating with domain experts throughout the model development and evaluation process can provide valuable insights into the clinical relevance and practical implications of the model's performance. Their feedback can guide the refinement of model architectures, hyperparameters and training strategies, ultimately leading to models that are better aligned with real-world clinical scenarios and decision-making processes.

Moreover, it's crucial to customize the augmentation strategies and evaluation methods according to the specific needs of each use case. For example, in medical imaging applications, the tolerance for missed detections may vary depending on the clinical situation.

For scenarios where missed diagnoses could have severe consequences, such as detecting life-threatening conditions, augmentation strategies should prioritize minimizing missed detections, even if it means more false alarms. This approach ensures that potential cases are not overlooked, enabling timely interventions and potentially saving lives.

On the other hand, in scenarios where false alarms may lead to unnecessary follow-up procedures or patient anxiety, augmentation strategies should balance minimizing missed detections and false alarms, optimizing for overall diagnostic accuracy and reducing the burden on healthcare resources.

Consequently, evaluation methods must also be carefully chosen and tailored to

align with the specific clinical priorities of each use case.

By customizing augmentation strategies and evaluation methods to the unique requirements of each imaging application, researchers can develop Mask R-CNN models that are optimized for the specific context, enhancing their practical value and potential impact on defect recognition.

The adoption of a pretrained neural network, such as ResNet-50, significantly streamlines the training process, leveraging the knowledge learned from large-scale datasets and reducing the computational resources required for model training. However, our observation of a plateau in loss reduction after Epoch 10 suggests potential implications for future experiments. This finding warrants further investigation into optimal training strategies and model convergence to ensure that the models are fully leveraging the available data and computational resources.

These findings highlight the multifaceted considerations involved in optimizing the performance of Mask R-CNN models and underscore the need for tailored approaches in data augmentation, dataset labeling, and model training in medical imaging applications.

7.2 Implications of the Results for Practice and Future Research

These findings have significant implications for both practical application and future research. Practically, the superiority of GAN-generated images suggests a promising approach for improving the diagnostic capabilities of Mask R-CNN models in skin cancer detection and other imaging tasks. By leveraging GANs or similar techniques, practitioners can enhance dataset diversity, thus improving model robustness and generalization.

Furthermore, this thesis underscores the importance of accurate labeling and the need for standardized protocols to minimize bias in training datasets. Future research should explore alternative data augmentation methods and investigate the potential of transfer learning techniques beyond ResNet-50. Additionally, extending this research to other imaging domains could yield valuable insights and advance deep learning solutions in healthcare or industry.

7.3 Future Work

7.3.1 Model Interpretability

Enhancing the interpretability of deep learning models like Mask R-CNN will be crucial for seamlessly integrating them into practical applications. Despite their promising performance in medical imaging tasks, these models' complex decision-making processes often lack transparency, hindering their adoption by experts. Incorporating interpretability techniques such as attention mechanisms and salience maps into the model architectures can provide valuable insights into how the models arrive at their predictions. By visualizing the features and regions of interest

driving the models' decisions, experts can validate the recommendations, identify potential biases or errors, and ultimately build trust in the models' outputs.

Fostering close collaboration between AI researchers and domain professionals will be essential for developing interpretable models tailored to specific workflows. This collaborative approach can bridge the gap between technical model development and practical requirements, ensuring that the models align with the needs and expectations of end-users. Through active engagement and knowledge exchange, AI researchers can gain a better understanding of the examined use-case, while professionals can provide valuable feedback on the interpretability and usability of the models.

Additionally, efforts should be made to develop standardized interpretability metrics and guidelines specific to medical imaging tasks. These metrics can quantify the degree of interpretability and transparency of the models, facilitating objective comparisons and enabling the selection of the most appropriate models for clinical use. By establishing consistent interpretability standards, the healthcare industry can promote the responsible development and deployment of deep learning models, ensuring patient safety and ethical decision-making.

However, the need for acceptance and accuracy extends beyond medical use cases. In industrial testing scenarios, for instance, the lack of official standards for automated computer vision applications poses challenges in integrating them into quality processes.

In essence, by prioritizing interpretability, fostering interdisciplinary collaboration, and establishing standardized interpretability metrics, the medical imaging community can pave the way for the successful integration of deep learning models into clinical practice, ultimately enhancing patient care and driving innovation in healthcare.

7.3.2 Workflow adaptability

The autonomous generation of synthetic data and training of Mask R-CNN models is a significant achievement in defect recognition research. Moving forward, it's crucial to validate the effectiveness of this approach across a wide range of datasets and defect recognition scenarios. The developed workflow's ability to translate defect recognition into a foreground/background scene opens up numerous potential applications in various industries, healthcare, and environmental domains.

To ensure widespread adoption and success, researchers should focus on extensive validation by testing the workflow on diverse datasets and scenarios. This will assess its adaptability, robustness, and performance under varying conditions, providing valuable insights into strengths, limitations, and areas for improvement.

Collaborative efforts among researchers are essential for optimizing the workflow. This includes exploring optimal data generation methods, determining the ideal synthetic data volume, and optimizing computational resource utilization. By leveraging diverse expertise and perspectives, more efficient and effective imple-

mentations can be achieved.

Establishing standardized benchmarking protocols and metrics is crucial for objectively evaluating and comparing the performance of different defect recognition approaches. This will facilitate the selection of appropriate methods for specific use cases and ensure consistent and transparent reporting of results.

Collaborating with domain experts from various industries, healthcare, and environmental sectors is vital for tailoring the workflow to their specific requirements and constraints. This will ensure that the approach addresses real-world challenges and integrates seamlessly into existing workflows and infrastructures.

Additionally, researchers should develop strategies to overcome potential computational limitations, such as leveraging distributed computing, optimizing resource allocation, or exploring efficient model architectures. This will enable the deployment of the approach in resource-constrained environments or for large-scale applications.

7.3.3 Optimizing Model Convergence Strategies

It may be beneficial to analyze the model's performance on a held-out validation set to assess whether the plateau in loss reduction corresponds to a plateau in performance metrics or if the model continues to improve despite the stagnant loss. This analysis could inform the decision to continue training beyond the observed plateau or to explore alternative architectures or training strategies.

Evaluating the model's performance on a separate validation set is crucial because the training loss alone may not provide a complete picture of the model's generalization capabilities. In some cases, the training loss may plateau, but the model's performance on unseen data could continue to improve, indicating that further training is beneficial.

Conversely, if the plateau in training loss corresponds to a plateau in performance metrics on the validation set, it may signal that the model has reached its maximum potential with the current architecture and training approach. In such cases, continuing to train beyond the observed plateau may not yield significant improvements and could even lead to overfitting.

By monitoring both the training loss and validation set performance, researchers can make informed decisions on whether to continue training, stop training, or explore alternative approaches. For instance, if the validation set performance continues to improve despite the stagnant training loss, it may be worthwhile to continue training for additional epochs or adjust the learning rate schedule to facilitate further learning.

7.3.4 Unlocking Multi-Class Capabilities

In the current scenario, the workflow is focused on a single-class use case, specifically detecting cancer cells without distinguishing between different types of cancer.

However, an important avenue for future research lies in extending the workflow to handle multiple classes.

Researchers could explore how the workflow can be adapted to classify different types of cancer or other abnormalities. For instance, in medical diagnostics, the ability to differentiate between various cancer types could significantly impact treatment decisions and patient outcomes. Similarly, in industrial applications such as surface inspection, distinguishing between different types of defects may inform maintenance or quality control procedures.

Extending the workflow to handle multiple classes presents both challenges and opportunities. Researchers may need access to more diverse and labeled datasets representing different classes, as well as robust machine learning techniques capable of handling multi-class classification tasks. Additionally, addressing potential class imbalances and computational complexities will be essential for the successful implementation of the workflow in real-world scenarios.

Despite these challenges, the benefits of multi-class classification are considerable. By accurately identifying and classifying different types of abnormalities or defects, the workflow can enhance diagnostic accuracy, enable targeted interventions, and ultimately improve outcomes across various domains.

7.3.5 Increase diversity of synthetic dataset

During this investigation, images generated by the cGAN exhibited a noticeable lack of diversity. To advance the effectiveness of synthetic data generation, future research should focus on optimizing the cGAN training process to produce a more diverse range of images.

This optimization effort should encompass various aspects, including the representation of different types of cancer cells and the inclusion of diverse skin color types in the training data. By ensuring a more comprehensive representation of the target domain, the cGAN can generate synthetic images that better reflect real-world variability, enhancing the robustness and generalizability of the trained models.

7.3.6 Alternatives to GANs

In future research, investigating different generative models offers a promising path to improve the performance of Mask R-CNN models in medical imaging. While GANs are effective in generating realistic images for augmentation, alternative models present unique opportunities for enhancement.

Variational Autoencoders, for instance, employ a probabilistic framework to learn latent representations of data. This method enables more controlled and interpretable generation of augmented images, potentially improving diversity and quality. Similarly, flow-based generative models provide efficient generation of diverse and high-quality samples by establishing invertible mappings between data space and latent space[18].

By exploring these alternative generative models, future research can expand the repertoire of augmentation techniques for Mask R-CNN models. This exploration has the potential to contribute to the development of more robust and generalizable models for medical image analysis, benefiting diagnostic accuracy and patient care.

References

[1] Samah AbuSalim et al. „Data Augmentation on Intra-Oral Images Using Image Manipulation Techniques“. In: *2022 International Conference on Digital Transformation and Intelligence (ICDI)*. 2022, pp. 117–120. DOI: 10.1109/ICDI57181.2022.10007158.

[2] Md. Shariful Alam, Dadong Wang, and Arcot Sowmya. „Image data augmentation for improving performance of deep learning-based model in pathological lung segmentation“. In: *2021 Digital Image Computing: Techniques and Applications (DICTA)*. 2021, pp. 1–5. DOI: 10.1109/DICTA52665.2021.9647209.

[3] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. „Understanding of a convolutional neural network“. In: *2017 International Conference on Engineering and Technology (ICET)*. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[4] Layth Alzubaidi, Jun Zhang, Ali J. Humaidi, et al. „Review of deep learning: concepts, CNN architectures, challenges, applications, future directions“. In: *J Big Data* (2021). URL: https://doi.org/10.1186/s40537-021-00444-8.

[5] Talha Azfar et al. *Deep Learning based Computer Vision Methods for Complex Traffic Environments Perception: A Review*. 2022. arXiv: 2211.05120 [cs.CV].

[6] Dan Becker. *Rectified Linear Units (ReLU) in Deep Learning*. https://www.kaggle.com/code/dansbecker/rectified-linear-units-relu-in-deep-learning. Accessed: 22 March 2024. 2018.

[7] S. Bhaggiaraj et al. „Deep Learning Based Self Driving Cars Using Computer Vision“. In: *2023 International Conference on Networking and Communications (ICNWC)*. 2023. DOI: 10.1109/ICNWC57852.2023.10127448.

[8] Andrew Tzer-Yeu Chen, Morteza Biglari-Abhari, and Kevin I-Kai Wang. „Trusting the Computer in Computer Vision: A Privacy-Affirming Framework“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017. DOI: 10.1109/CVPRW.2017.178.

[9] Vinicius Luis Trevisan De Souza, Bruno Augusto Dorta Marques, and João Paulo Gois. „Fundamentals and Challenges of Generative Adversarial Networks for Image-based Applications“. In: *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. Vol. 1. 2022, pp. 308–313. DOI: 10.1109/SIBGRAPI55357.2022.9991776.

[10] Ivan Dokmanic et al. „Euclidean Distance Matrices: Essential theory, algorithms, and applications“. In: *IEEE Signal Processing Magazine* 32.6 (Nov. 2015). ISSN: 1053-5888. DOI: 10.1109/msp.2015.2398954. URL: http://dx.doi.org/10.1109/MSP.2015.2398954.

[11] Pavel Hamet and Johanne Tremblay. „Artificial intelligence in medicine“. In: *Metabolism* 69 (2017). Insights Into the Future of Medicine: Technologies, Concepts, and Integration, S36–S40. ISSN: 0026-0495. DOI: https://doi.org/10.1016/j.metabol.2017.01.011. URL: https://www.sciencedirect.com/science/article/pii/S002604951730015X.

[12] Hongjie He et al. „Mask R-CNN based automated identification and extraction of oil well sites“. In: *International Journal of Applied Earth Observation and Geoinformation* 112 (Aug. 2022), p. 102875. DOI: 10.1016/j.jag.2022.102875.

[13] Kaiming He et al. „Mask R-CNN“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, p. 1. DOI: 10.1109/ICCV.2017.322.

[14] Kaiming He et al. „Mask R-CNN“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, p. 2. DOI: 10.1109/ICCV.2017.322.

[15] Kaiming He et al. „Mask R-CNN“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, p. 3. DOI: 10.1109/ICCV.2017.322.

[16] Kaiming He et al. „Mask R-CNN“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, p. 5. DOI: 10.1109/ICCV.2017.322.

[17] Suzana Herculano-Houze. *The Human Brain in Numbers: A Linearly Scaled-up Primate Brain*. 2009. ront Hum Neurosci: 19915731.

[18] Ashhadul Islam and Samir Brahim Belhaouari. „Fast and Efficient Image Generation Using Variational Autoencoders and K-Nearest Neighbor OveRsampling Approach“. In: *IEEE Access* 11 (2023), pp. 28416–28426. DOI: 10.1109/ACCESS.2023.3259236.

[19] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. 2018. arXiv: 1611.07004 [cs.CV].

[20] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV].

[21] Zhelin Liu et al. „Recent Advances of Generative Adversarial Networks“. In: *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*. 2022, pp. 558–562. DOI: 10.1109/ICETCI55101.2022.9832194.

[22] K. Mader. *Skin Cancer MNIST: HAM10000*. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000. Accessed: 15.01.2024. 2018.

[23] McKinsey. *The state of AI in 2023: Generative AI's breakout year*. Aug. 2023. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year (visited on 03/19/2024).

[24] Agnieszka Mikołajczyk and Michał Grochowski. „Data augmentation for improving deep learning in image classification problem“. In: *2018 International Interdisciplinary PhD Workshop (IIPhDW)*. 2018, pp. 117–122. DOI: 10.1109/IIPHDW.2018.8388338.

[25] Keiron O'Shea and Ryan Nash. „An Introduction to Convolutional Neural Networks“. In: *CoRR* abs/1511.08458 (2015). arXiv: 1511.08458. URL: http://arxiv.org/abs/1511.08458.

[26] José Crossa Osval Antonio Montesinos López Abelardo Montesinos López. *Fundamentals Artificial Neural Networks and Deep Learning*. Springer, 2021, p. 380.

[27] José Crossa Osval Antonio Montesinos López Abelardo Montesinos López. *Fundamentals of Artificial Neural Networks and Deep Learning*. Springer, 2021, p. 379.

[28] José Crossa Osval Antonio Montesinos López Abelardo Montesinos López. *Fundamentals of Artificial Neural Networks and Deep Learning*. Springer, 2021, p. 381.

[29] José Crossa Osval Antonio Montesinos López Abelardo Montesinos López. *Fundamentals of Artificial Neural Networks and Deep Learning*. Springer, 2021, p. 383.

[30] José Crossa Osval Antonio Montesinos López Abelardo Montesinos López. *Fundamentals of Artificial Neural Networks and Deep Learning*. Springer, 2021, p. 384.

[31] José Crossa Osval Antonio Montesinos López Abelardo Montesinos López. *Fundamentals of Artificial Neural Networks and Deep Learning*. Springer, 2021, p. 385.

[32] Deepak Pathak et al. *Context Encoders: Feature Learning by Inpainting*. 2016. arXiv: 1604.07379 [cs.CV].

[33] Luis Perez and Jason Wang. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. 2017. arXiv: 1712.04621 [cs.CV].

[34] P. Jonathan Phillips, R. Michael McCabe, and Rama Chellappa. „Biometric image processing and recognition“. In: *9th European Signal Processing Conference (EUSIPCO 1998)*. 1998.

[35] Eko Prasetyo, Nanik Suciati, and Chastine Fatichah. „A Comparison of YOLO and Mask R-CNN for Segmenting Head and Tail of Fish“. In: *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*. DOI: 10.1109/ICICoS51170.2020.9299024.

[36] *PyTorch: An open source deep learning platform*. https://pytorch.org/docs/stable/index.html. Accessed: March 20, 2024.

[37] Qiong Qiao. „Image Processing Technology Based on Machine Learning“. In: *IEEE Consumer Electronics Magazine* (2022), pp. 1–1. DOI: 10.1109/MCE.2022.3150659.

[38] Xiaoli Qin, Francis M. Bui, and Ha H. Nguyen. „Learning from an Imbalanced and Limited Dataset and an Application to Medical Imaging“. In: *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. 2019, pp. 1–6. DOI: 10.1109/PACRIM47961.2019.8985057.

[39] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV].

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[41] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: 1409.0575 [cs.CV].

[42] I.H. Sarker. *Machine Learning: Algorithms, Real-World Applications and Research Directions*. 2021. eprint: https://doi.org/10.1007/s42979-021-00592-x.

[43] Viktor Seib, Benjamin Lange, and Stefan Wirtz. „Mixing Real and Synthetic Data to Enhance Neural Network Training - A Review of Current Approaches“. In: *CoRR* abs/2007.08781 (2020). arXiv: 2007.08781. URL: https://arxiv.org/abs/2007.08781.

[44] Andrzej Sioma. „Vision System in Product Quality Control Systems“. In: *Applied Sciences* 13.2 (2023). ISSN: 2076-3417. URL: https://www.mdpi.com/2076-3417/13/2/751.

[45] *Supervisely*. https://supervisely.com/. 15.01. 2024.

[46] Pauli Virtanen et al. „SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python“. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[47] Tobias Wolf. *SkinCancerworkflow*. https://github.com/tobwol/SkinCancerworkflow. Accessed: 2024-05-26. 2024.

[48] Matthew D Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*. 2013. arXiv: 1311.2901 [cs.CV].

[49] Jiajun Zhang, Georgina Cosma, and Jason Watkins. „Image Enhanced Mask R-CNN: A Deep Learning Pipeline with New Evaluation Measures for Wind Turbine Blade Defect Detection and Classification“. In: *Journal of Imaging* 7.3 (2021). ISSN: 2313-433X. DOI: 10.3390/jimaging7030046. URL: https://www.mdpi.com/2313-433X/7/3/46.

[50] Yanying Zhang and Xiaolin Zheng. „Development of Image Processing Based on Deep Learning Algorithm“. In: *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*. 2022. DOI: 10.1109/IPEC54454.2022.9777479.