HAW HAMBURG

BACHELOR THESIS
Niklas Hoefflin

Effects of Image Resolution on Skin Lesions Classification with Residual Neural Networks

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Niklas Hoefflin

Effects of Image Resolution on Skin Lesions Classification with Residual Neural Networks

Bachelor thesis submitted for examination in Bachelor´s degree
in the study course *Bachelor of Science Angewandte Informatik*
at the Department Computer Science
at the Faculty of Engineering and Computer Science
at University of Applied Science Hamburg

Supervisor: Prof. Dr. Kai von Luck
Supervisor: Dr. Jan Schwarzer

Submitted on: July 2, 2024

Niklas Hoefflin

Title of Thesis

Effects of Image Resolution on Skin Lesions Classification with Residual Neural Networks

Keywords

Medical Imaging, Dermatology, Skin Lesions, Image Resolution, Computer Vision, Residual Neural Network

Abstract

Accurate detection of tumorous skin lesions is crucial for the early diagnosis and treatment of skin cancer, significantly impacting patient health outcomes. High-resolution images of skin lesions are vital for enhancing the classification performance of deep learning models. These detailed images provide critical visual information, enabling more precise identification and differentiation of malignant and benign lesions, ultimately improving patient care. However, the effect of image resolution, especially for images with aspect ratios above 450×450 pixels, remains unclear. This thesis investigated the influence of image resolution on the classification performance of skin lesion images using the well-established Residual Neural Network (ResNet) architecture. Specifically, the study examined three different resolutions: 300×225, 450×338, and 600×450 pixels. Extensive experiments were conducted and evaluated using the Area Under The Receiver Operating Characteristic Curve (AUROC) metric. The results showed a positive correlation between increasing image resolution and increasing AUROC scores across all classes. Although the degree of performance improvement seemed to be class-dependent, the one standard deviation variability in terms of the AUROC score decreased with increasing resolutions, with the most notable reduction observed at 600×450 pixels.

Niklas Hoefflin

Thema der Arbeit

Auswirkungen der Bildauflösung auf die Klassifizierung von Hautläsionen mit Residual Neural Networks

Stichworte

Medizinische Bildgebung, Dermatologie, Hautläsionen, Bildauflösung, Bilderkennung, Residual Neural Network

Kurzzusammenfassung

Die genaue Erkennung tumoröser Hautläsionen stellt einen entscheidenden Faktor bei der frühzeitigen Diagnose und Behandlung von Hautkrebs dar und hat signifikante Auswirkungen auf die gesundheitliche Verfassung der Patienten. Hochauflösende Bilder von Hautläsionen sind von entscheidender Bedeutung für die Verbesserung der Klassifizierungsleistung von Deep-Learning-Modellen. Die Bereitstellung detaillierter Bilder ermöglicht eine präzisere Identifizierung und Unterscheidung von bösartigen und gutartigen Läsionen, was letztlich zu einer Verbesserung der Patientenversorgung führt. Die Auswirkungen der Bildauflösung, insbesondere bei Bildern mit einem Seitenverhältnis von mehr als 450×450 Pixeln, sind jedoch noch unklar. In dieser Arbeit wurde der Einfluss der Bildauflösung auf die Klassifizierungsleistung von Hautläsionsbildern unter Verwendung der bewährten Residual Neural Network (ResNet)-Architektur untersucht. Konkret wurden drei verschiedene Auflösungen untersucht: 300×225, 450×338 und 600×450 Pixel. Es wurden umfangreiche Experimente durchgeführt, die anhand der Area Under The Receiver Operating Characteristic Curve (AUROC)-Metrik ausgewertet wurden. Die Ergebnisse zeigten eine positive Korrelation zwischen steigender Bildauflösung und steigenden AUROC-Werten in allen Klassen. Obwohl der Grad der Leistungsverbesserung klassenabhängig zu sein schien, nahm die Streuung von einer Standardabweichung in Bezug auf den AUROC-Score mit zunehmender Auflösung ab, wobei die bemerkenswerteste Verringerung bei 600×450 Pixeln beobachtet wurde.

Acknowledgments

First of all, I would like to express my gratitude to my supervisors, Prof. Dr. Kai von Luck and Dr. Jan Schwarzer, for their guidance, feedback, and encouragement throughout the entire research process. Their expertise was invaluable and greatly contributed to the completion of this thesis.

I would also like to thank my proofreaders, Tim Spulak, Nikolaus König, and Kacem Mbarek for their valuable feedback and suggestions on the preliminary drafts of this thesis.

Finally, I am deeply grateful to my family for their years of personal and financial support, without which this work, among other things, would not have been possible. Special thanks go to my grandmother, who introduced me to computer science at an early age and sparked my interest in this field.

Contents

List of Figures

List of Tables

Abbreviations

AK	Actinic Keratosis.
AUROC	Area Under The Receiver Operating Characteristic Curve.
BCC	Basal Cell Carcinoma.
BKL	Benign Keratosis (Solar Lentigo/ Seborrheic Keratosis/ Lichen Planus-like Keratosis).
BMCA	Balanced Multi-Class Accuracy.
CAD	Computer-aided Diagnosis.
CNN	Convolutional Neural Network.
CSTI	Creative Space For Technical Innovations.
DCA	Dark Corner Artifact.
DCGAN	Deep Convolutional Generative Adversarial Network.
DCNN	Deep Convolutional Neural Network.
DDIM	Denoising Di□usion Implicit Model.
DF	Dermatofibroma.
FN	False Negative.
FP	False Positive.
FPR	False Positive Rate.
GAN	Generative Adversarial Network.
Grad-CAM	Gradient-weighted Class Activation Mapping.
IECET	Illumination Equalization Method Based On A Counter Exponential Transform.

ISIC	International Skin Imaging Collaboration.
MEL	Melanoma.
MSC	Melanoma Skin Cancer.
MSM-CNN	Multiscale Multi-CNN.
NMSC	Non-Melanoma Skin Cancer.
NST	Neural Style Transfer.
NV	Melanocytic Nevus.
OpenCV	Open Source Computer Vision Library.
ReLU	Rectified Linear Unit.
ResNet	Residual Neural Network.
ROC	Receiver Operating Characteristic.
SCC	Squamous Cell Carcinoma.
SE	Structuring Element.
TN	True Negative.
TP	True Positive.
TPR	True Positive Rate.
VASC	Vascular Lesion.

1 Introduction

Skin cancer is one of the most commonly diagnosed types of cancer worldwide [75], posing a significant health concern across populations. In 2020, an estimated 1.5 million new cases of skin cancer occurred, while malignant melanoma accounts for approximately 20% of all skin cancer diagnoses, with an estimated 325,000 cases and 57,000 deaths worldwide [7]. Generally, skin cancer is commonly categorized into two types: Melanoma Skin Cancer (MSC) and Non-Melanoma Skin Cancer (NMSC) [31, 56]. NMSC is more common and less aggressive, while MSC poses a greater threat and is more lethal due to its propensity for metastasis [18]. Early diagnosis significantly improves survival rates [59, 64]. However, in advanced stages with metastasis, the 5-year survival rate drops to only between 15–20% [50].

Given the significance of early detection, dermatologists employ a range of techniques to identify skin cancer at an early stage. Dermoscopy is a non-invasive procedure that employs a dermatoscope to identify skin abnormalities and reveal structures and patterns below the surface. Algorithmic methods, including the Asymmetry, Border, Color, Diameter rule, the Menzies method, the three-point checklist, the seven-point checklist, and the Color, Architecture, Symmetry, Homogeneity algorithm, are employed to distinguish between NMSC and MSC based on visual characteristics [42, 69]. Another diagnostic procedure is biopsy, an invasive technique in which a small sample or the entire suspicious skin lesion is removed using methods such as punch, shave, saucerization, or excisional biopsy. The sample is then subjected to histopathological analysis under a microscope for an accurate diagnosis of the disease type [49].

Although these methods are well-established, manual interpretations prove to be time-consuming and costly [30, 79]. Moreover, distinguishing between MSC and NMSC requires considerable expertise and is difficult even for experienced dermatologists due to the high level of inter-class similarity and intra-class variations [35, 54, 79]. Computer-aided Diagnosis (CAD) systems, based on Convolutional Neural Networks (CNNs), have

shown promising results in both enhancing the e□ciency and accuracy of skin cancer diagnosis, closely aligning with dermatologists in experimental settings [20].

1.1 Research Aim

While significant advancements have been made in CAD, the quality of training images remains a critical factor influencing the performance of CNNs. Among the various aspects of image quality, resolution plays a pivotal role, particularly in the classification of medical images such as chest radiographs, endoscopic images, breast ultrasounds, and skin lesion images [38, 39, 55, 66, 67]. In two studies [38, 39], Mahbod et al. investigated the impact of varying image resolutions on the performance of pre-trained CNNs in the classification of skin lesions. In one study [38], the influence of image resolutions up to 450×450 pixels was examined while maintaining the aspect ratio. However, in another study [39], the authors used resolutions up to 768×768 pixels, although without maintaining the aspect ratio.

This study aims to address the existing knowledge gap regarding the e□ect of image resolution on the performance of CNNs for resolutions above 450×450 pixels while maintaining aspect ratio. Utilizing the International Skin Imaging Collaboration (ISIC) 2019 challenge dataset [14, 16, 68], the objective of this study is to provide a more profound understanding of the optimal image resolution for enhancing the performance of CNNs in the classification of skin lesions.

1.2 Outline

This thesis is structured into five chapters. Following the introduction, Chapter 2 provides a theoretical foundation and reviews related work, followed by the research question. Chapter 3 outlines the methodologies employed to address the research question, including dataset preparation and their subsets, pre-processing, data augmentation techniques, and the design of model architectures, as well as the training procedures for the experiments. Chapter 4 evaluates the results and discusses their e□ects and limitations. Finally, Chapter 5 o□ers a conclusion that summarizes the findings and suggests directions for future research.

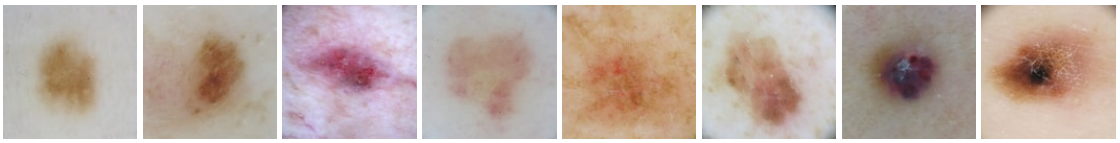
2 Related Work

This chapter provides a theoretical foundation, reviews related work, introduces the problem of image resolution in medical imaging, and presents the research question that this study aims to answer. In Section 2.1, the ISIC 2019 dataset utilized in this study is introduced. Section 2.2 discusses the challenges associated with metrics in imbalanced datasets, highlighting the Area Under The Receiver Operating Characteristic Curve (AUROC) metric used in this study. Furthermore, Section 2.3 provides an introduction to data augmentation methods along with relevant studies, while Section 2.4 introduces various pre-processing techniques. Section 2.5 introduces residual neural networks and strategies to enhance model performance through transfer learning and fine-tuning. Lastly, Section 2.6 presents the issue of image resolution in medical imaging with relevant studies before formulating the research question in Section 2.7.

2.1 ISIC 2019 Dataset

The ISIC 2019 challenge dataset comprises 33,569 dermoscopic images of skin lesions, ranging in resolutions from 600×450 to 1024×1024 pixels. These images originate from three di□erent datasets: HAM10000 [68], BCN20000 [16], and MSK [14]. Participants in the ISIC 2019 challenge employed this dataset to develop algorithms capable of distinguishing between di□erent types of skin lesions. The images are taken from multiple body sites and are each associated with clinical metadata reviewed by recognized melanoma experts. However, this metadata is not utilized in this study. The dataset has been divided into two subsets: 25,331 training images and 8,238 test images. Despite the large number of images, the dataset exhibits a highly imbalanced class distribution, which can negatively impact the accuracy of deep learning models [24]. The exact class distribution is further discussed in Section 3.1. The images are categorized into nine distinct diagnostic classes, with the ninth class representing an unknown category present only in the test images.

The eight known classes shown in Figure 2.1 include: Melanocytic Nevus (NV), Melanoma (MEL), Basal Cell Carcinoma (BCC), Benign Keratosis (Solar Lentigo/ Seborrheic Keratosis/ Lichen planus-like Keratosis) (BKL), Actinic Keratosis (AK), Squamous Cell Carcinoma (SCC), Vascular Lesion (VASC), and Dermatofibroma (DF).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 2.1: Skin lesions present in the ISIC 2019 challenge dataset: (a) NV skin lesion, (b) MEL skin lesion, (c) BCC skin lesion, (d) BKL skin lesion, (e) AK skin lesion, (f) SCC skin lesion, (g) VASC skin lesion, and (h) DF skin lesion.

The HAM10000 dataset consists of 10,015 dermoscopic images collected over a period of 20 years from the Department of Dermatology at the Medical University of Vienna, Austria, and from the skin cancer practice of Cli Rosendahl in Queensland, Australia. More than 50% of these images were confirmed pathologically, with the remaining images verified through follow-up, expert consensus, or in vivo confocal microscopy. The BCN20000 dataset includes 19,424 images, many of which pose significant classification challenges due to factors such as lighting conditions, anatomical location, and other confounding variables. These images were collected over more than 16 years by the Department of Dermatology at the Hospital Clínic de Barcelona. Lastly, the MSK dataset comprises 2,000 images gathered from the Memorial Sloan Kettering Cancer Center.

2.2 Performance Evaluation

Performance evaluation is crucial for assessing the e ectiveness of a model trained on data for classification tasks. Various metrics can be employed to assess the correctness of classification. One of the most frequently employed metrics in the domains of machine learning and statistics is accuracy [28, 44]. Accuracy measures the ratio of correctly classified instances out of the total instances in the dataset and is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.1}$$

where True Positive (TP) represents instances correctly predicted as positive, True Negative (TN) represents instances correctly predicted as negative, False Positive (FP) represents instances incorrectly predicted as positive, and False Negative (FN) represents instances incorrectly predicted as negative. In medical diagnostics, where data is often imbalanced [21], the metric can be misleading due to its characteristic of only capturing the overall ratio of correctly classified instances without considering the class distribution. Consequently, a model might achieve high accuracy by predicting the majority class for every instance while performing poorly on the minority classes [28]. The application of the accuracy metric to imbalanced datasets is therefore not an e□ective approach [33], as the metric is unable to accurately reflect the true capabilities of the model in classifying samples in terms of class imbalance. This is of particular importance in the field of medical diagnostics, as it may result in the overlooking of conditions that can have a significant impact on patient health [33].

The Receiver Operating Characteristic (ROC) curve analysis is another method used in medical diagnostics to visually assess the performance of diagnostic tests or predictive models in distinguishing the presence or absence of a certain health condition [10, 46]. It visually represents the trade-o□ between the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis across each classification threshold [44]. TPR represents the ratio of actual positives correctly identified by the model and is defined as

$$TPR = \frac{TP}{TP + FN}. \tag{2.2}$$

Concurrently, FPR, which denotes the ratio of false positives within the negative class, is defined as

$$FPR = \frac{FP}{FP + TN}. \tag{2.3}$$

While the ROC curve provides a graphical representation of the performance across various classification thresholds, the AUROC metric quantifies the overall performance of a model across all thresholds in a single score. Unlike accuracy, it is not a□ected by imbalanced datasets [57]. AUROC represents the integral of the ROC curve over the range $[0, 1]$ and is defined as

$$\text{AUROC} = \int_0^1 \text{TPR}(FPR)\,\mathrm{d}FPR. \tag{2.4}$$

The integral is approximated using the trapezoidal rule, which involves dividing the area under the curve into small trapezoids and summing their areas. According to Nahm [46],

different interpretation thresholds, as shown in Figure 2.2, determine when a diagnostic technique is considered meaningful.

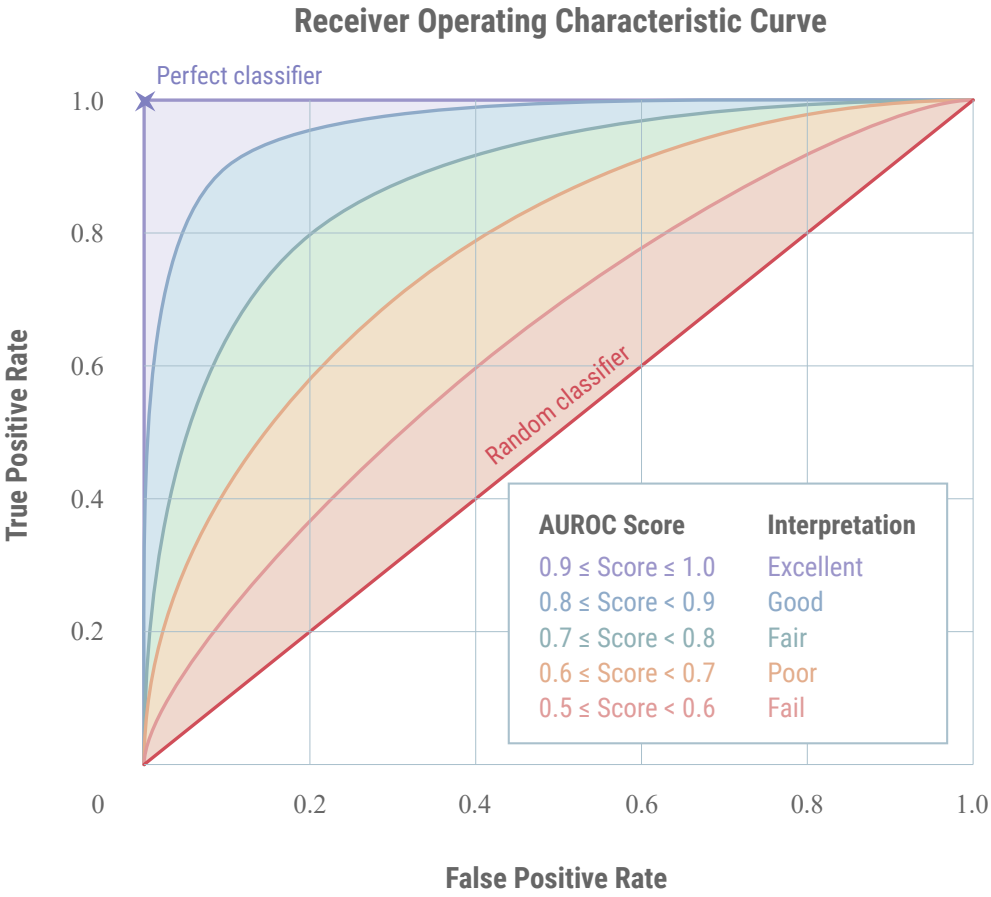


Figure 2.2: Interpretation of AUROC scores across various thresholds (based on Nahm [46], p. 6). A score closer to *1.0* indicates better model performance in distinguishing between classes. Scores around and up to *0.8* are generally still considered acceptable [45, 46]. Scores closer to *0.5* indicate performance similar to that of a random classifier. Scores below *0.5* are not meaningful.

2.3 Data Augmentation

The limited size and imbalanced class distribution within datasets are common challenges in deep learning, especially in medical imaging, often leading to overfitting [21, 22, 24, 32]. Larger datasets typically yield better results and improved generalization [61]. However, acquiring and annotating high-quality medical images for skin cancer detection is costly and time-consuming [30]. Creating these images requires expensive equipment, and accurate annotating depends on the expertise of dermatologists, who must discern inter-class similarities and intra-class variations in skin lesions [35].

To address these challenges, data augmentation techniques are employed to enhance the quantity and diversity of training data for deep learning models while reducing overfitting and improving generalization performance [21, 61]. These techniques encompass a variety of methods designed to create more robust and varied datasets. Geometric transformations [21, 32], such as flipping, cropping, rotation, and shifting, alter the spatial properties of images, while color transformations [21, 32], including adjustments in brightness, saturation, contrast, and hue, modify the visual characteristics.

In skin cancer detection, both o□ine and online augmentation methods are commonly used to expand the dataset. Minimizing false positive and false negative predictions is crucial, as misclassifications can have serious consequences for patient care [33]. Therefore, it is essential to evaluate whether o□ine or online augmentation is more suitable for a given problem.

In o□ine augmentation, images are augmented and stored prior to the training of deep learning models, reducing computational demand during training and thereby shortening training time compared to online augmentation [61]. This approach is particularly beneficial for imbalanced datasets, as it allows for more balanced class distributions through over- and undersampling techniques. Oversampling increases the instances of underrepresented classes to mitigate model bias toward the majority class, while undersampling removes instances from the majority class to achieve class balance [11, 24, 61]. For instance, Bozkurt [9] expanded the HAM10000 skin lesion dataset from 10,015 to 39,787 images using random rotations. This augmented dataset enabled an Inception-Residual Neural Network (ResNet)-v2 model to achieve an accuracy of 95.09%, compared to 83.59% on the original dataset, surpassing other architectures such as VGG16, VGG19, SqueezeNet, LeNet-5, AlexNet, and Deep Convolutional Neural Network (DCNN). Similarly, Ahmad et al. [3] used o□ine augmentation for the HAM10000 and ISIC 2018

[15] datasets, increasing each from 10,000 to 36,000 images through 90-degree rotations and flips. They applied transfer learning by fine-tuning pre-trained Xception and ShuffleNet models on the augmented data, extracting features from global average pooling layers, and using an Improved Bat Optimization Algorithm for feature selection and classification. This method, supported by Gradient-weighted Class Activation Mapping (Grad-CAM) for interpretability, achieved high accuracy on both datasets. Furthermore, Ali et al. [4] developed a DCNN for the classification of benign and malignant skin lesions in the HAM10000 dataset. By applying mean filtering, normalization, and augmentation techniques such as rotation, cropping, flipping, and color-shifting, the model demonstrated superior performance compared to pre-trained architectures, including AlexNet, ResNet, VGG-16, DenseNet, and MobileNet, achieving accuracies of 90.16% and 91.43% for different training set splits.

In contrast to offline augmentation, where data is pre-augmented and stored before training, online augmentation dynamically and randomly augments data during training [61]. Before each epoch begins, various transformations are applied to the images, ensuring that the model consistently trains on unique data variations. This method, though more resource-intensive due to on-the-fly augmentation, enhances the model's generalization capabilities by exposing it to new data variations constantly. Nie et al. [48] leveraged online augmentation for skin lesion classification by introducing a hybrid CNN-Transformer model that combines ResNet-50 and Vision Transformer architectures. They trained this model on the ISIC 2018 dataset using online augmentation techniques such as flipping, rotating, scaling, cropping, affine transformations, and brightness and contrast adjustments. Their experiments demonstrated that the hybrid model, when compared to the baseline ResNet-50, consistently achieved higher accuracy, particularly with the focal loss function, reaching an overall accuracy of 89.48%. He et al. [27] proposed the Deep Metric Attention Learning CNN for skin lesion classification, utilizing online augmentation techniques like random rotations and flips to mitigate overfitting on the ISIC 2016 [23] and ISIC 2017 [14] datasets. This model outperformed other methods and showed strong generalization capabilities on the PH^2 dataset [41].

In addition to conventional augmentation techniques, advanced techniques such as Generative Adversarial Networks (GANs), Neural Style Transfer (NST), diffusion models, and meta-learning methods are employed to generate synthetic data, enhancing dataset diversity and potentially improving model performance. Abhishek and Hamarneh [1] introduced Mask2Lesion, a GAN-based approach that generates synthetic skin lesion images from segmentation masks. This method improved segmentation accuracy on the

ISIC 2017 dataset, outperforming classical augmentation methods, particularly when used in conjunction with them. Rezk et al. [53] addressed the issue of diversity in skin lesion datasets by generating synthetic images of darker skin tones using NST and deep blending. They trained ResNet-50 models on these synthetic images from the DermNet NZ[1] and ISIC 2018 JID Editorial Images[2] datasets and evaluated them on the Dermatology Atlas dataset. A visual Turing test and disease identification test validated the realism of NST-generated images, which significantly improved diagnostic accuracy and model performance. Eggert [19] developed a software solution utilizing a Deep Convolutional Generative Adversarial Network (DCGAN) and a Denoising Di□usion Implicit Model (DDIM) to generate synthetic photorealistic images from user-provided datasets. His experiments showed that the DDIM, particularly when combined with data augmentation techniques, outperformed the DCGAN in generating synthetic images. These synthetic images notably enhanced the performance of an object detector when used for training, highlighting the practical utility of synthetic data generation methods in computer vision tasks.

2.4 Pre-processing

Pre-processing is a crucial step in preparing images of skin lesions for training CNNs. E□ective pre-processing can significantly improve the quality of input data, leading to better model performance [36]. Skin lesion images often contain artifacts such as hair and Dark Corner Artifacts (DCAs), which can adversely a□ect performance if not addressed [43, 63, 72].

Alphonse et al. [5] addressed hair artifacts using the DullRazor algorithm, which enhances image clarity and model performance. They combined the Sobel Directional Pattern feature extraction method with a stacked Restricted Boltzmann Machine classifier, achieving high accuracy in melanoma prediction across multiple datasets, including PH[2], ISIC 2016, ISIC 2017, DermNet NZ, and DermIS[3]. Pewton and Yap [51] introduced an automated DCA detection and removal method evaluated on the Dermofit image library[4].

[1] https://www.dermnetnz.org/ (Accessed 05/23/24)
[2] https://api.isic-archive.com/collections/62/ (Accessed 05/23/24)
[3] https://www.dermis.net/dermisroot/en/home/index.htm (Accessed 05/23/24)
[4] https://licensing.edinburgh-innovations.ed.ac.uk/product/dermofit-image-library (Accessed 05/23/24)

Although the deep learning performance showed marginal differences, Grad-CAM visualizations indicated that removing DCAs shifted network activations more accurately onto the skin lesions.

Non-uniform lighting conditions in skin lesion images can alter crucial regions, leading to over- or under-lighting that negatively affects the performance of the model. Venugopal et al. [70] proposed an automated illumination equalization method based on a counter exponential transform (IECET) to address uneven lighting. Their findings indicate that applying IECET improved segmentation accuracy across models such as DTP-Net, DeepLabV3+, FCN, and U-Net. In another study, Afza et al. [2] developed a hybrid contrast stretching technique to enhance contrast in skin lesion images from the HAM10000 and ISIC 2018 datasets. They utilized NasNet for feature extraction, followed by feature selection using whale optimization and entropy-mutual information. The optimal features were fused using modified canonical correlation analysis and classified using an extreme learning machine. This automated deep learning-based framework demonstrated high accuracy, outperforming other methods on the datasets used. Alshahrani et al. [6] applied several pre-processing techniques to the ISIC 2019 dataset, including average filtering to reduce noise and smooth images. They also used contrast-limited adaptive histogram equalization to enhance the edges of skin lesions, revealing low-contrast areas. Their proposed Hybrid-CNN models achieved high accuracy in distinguishing skin cancers from other lesions.

Artifacts such as blood vessels, clinical markings, air bubbles, ruler marks, or size reference stickers are more challenging to remove due to their unpredictability and irregular and discontinuous nature [77]. Segmentation can significantly enhance pre-processing by isolating these artifacts from skin lesions, thereby improving classification accuracy [74]. Nida et al. [47] used morphological operations to remove hairs, blood vessels, and clinical marks from skin lesion images, further enhancing them with an unsharp filter. They trained a region-based CNN to localize melanoma-affected regions, which were then segmented using fuzzy C-means. This method demonstrated superior segmentation results compared to other techniques. Furthermore, Zafar et al. [78] proposed an automated skin lesion segmentation method combining U-Net and ResNet architectures, called Res-Unet. They addressed poor contrast issues through resizing and normalization and removed hairs using morphological operations, subsequently inpainting them with neighboring pixel values. The model, trained on the ISIC 2017 dataset and evaluated on the ISIC 2017 and PH^2 test sets, achieved results comparable to state-of-the-art techniques at the time.

2.5 Residual Neural Network

ResNets, introduced by He et al. [26] in 2015, are CNN architectures designed to address the vanishing gradient problem by incorporating residual blocks. The vanishing gradient problem occurs during the training of deep neural networks when the gradients used to update the weights during backpropagation become very small. This makes it difficult for the network to learn new features from the training data, leading to poor performance. A residual block shown in Figure 2.3 is a combination of convolutional layers and a shortcut connection that performs identity mapping by directly adding the input of the block to its output. This shortcut connection allows the gradients to skip one or more layers, which helps to maintain stronger gradient signals as they propagate backward through the network [26].

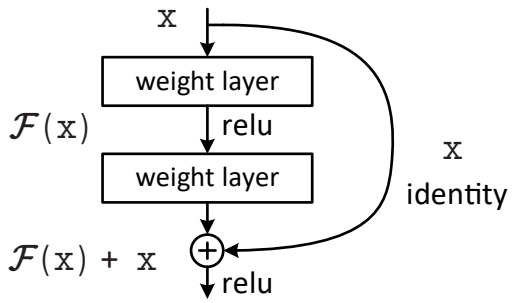


Figure 2.3: Residual Block with a shortcut connection (from He et al. [26] p. 2).

The ability of ResNet to train very deep networks effectively has made it a popular choice for a variety of applications, including image classification tasks such as autism detection in magnetic resonance imaging brain images [29], pneumonia detection in chest X-ray images [62], and colorectal cancer detection in colon gland images [58].

However, one common challenge in such tasks, particularly in specialized domains such as medical imaging, is the limited availability of labeled images. This lack of labeled data poses a significant hurdle for training robust skin lesion classification models [21]. One effective solution to this problem is the use of transfer learning. Transfer learning addresses this issue by leveraging models pre-trained on tasks within a related domain as a starting point for training models on similar tasks within the target domain, thereby improving learning performance, preventing overfitting, and reducing the amount of required labeled data in the target domain [12, 61, 79]. In transfer learning [12, 40], the head (fully connected layers) of a pre-trained model is often replaced with new layers.

Initially, the body (feature extractor) of the model is typically frozen, meaning that its weights are not updated during the training of the new head to better capture lower-level features in the new dataset learned from training on large datasets, such as ImageNet [17]. Another method related to transfer learning is fine-tuning [12, 40], where some or all layers of the body are unfrozen and trained further on the new dataset using lower learning rates. This step allows the pre-trained layers to adapt more closely to the new data to enhance the performance of the model without disrupting the previously learned features [12].

2.6 Effect of Image Resolutions in Medical Imaging

Training deep learning models with high-resolution images is more time-consuming and computationally demanding compared to using lower-resolution images [38, 39, 55]. However, reducing the resolution can lead to the loss of crucial medically relevant information, which can adversely affect the classification accuracy of deep learning models [38, 39, 55] and potentially have serious health implications for patients [67]. Thus, it is essential to find a balance between the optimal resolution for training deep learning models and achieving the best possible classification performance.

Sabottke and Spieler [55] studied the performance differences at various resolutions of chest X-ray images from the NIH ChestX-ray14 dataset [73], ranging from 32×32 to 600×600 pixels. They utilized pre-trained ResNet-34 and DenseNet-121 architectures from ImageNet. The highest AUROC scores for binary classification of different classes were observed at resolutions between 256×256 and 448×448 pixels. It was also found that some classes benefited more from higher image resolution than others, especially for classes that appeared more scattered in the images, as important information was lost by downscaling. Similarly, Thambawita et al. [67] analyzed the performance of pre-trained ResNet and DenseNet architectures on endoscopy images with different resolutions, ranging from 32×32 to 512×512 pixels, using the HyperKvasir dataset [8]. Results showed that increasing image resolution correlates with an increase in the Matthews Correlation Coefficient for all resolutions, with the highest score achieved at 512×512 pixels. The authors concluded that image resolution has a clear influence on the performance of deep learning models in the classification of endoscopy images. Tang et al. [66] investigated the effects of different image resolutions on ultrasound breast images collected from two hospitals, using MobileNet and DenseNet-121 architectures with resolutions ranging from 64×64

to 512×512 pixels. MobileNet achieved the highest AUROC score with a confidence interval of 95% at a resolution of 320×320 pixels, while DenseNet-121 achieved the highest AUROC at 448×448 pixels. Their results showed a high correlation between changes in image resolution and breast cancer diagnosis accuracy in ultrasound breast images.

According to Mahbod et al. [38, 39], there is an open question in the field of skin lesion classification regarding the optimal resizing factor for fine-tuning pre-trained CNNs, as downsampling may lead to a loss of useful medical information. In their study [38], they examined the effect of different image resolutions in skin lesion classification using pre-trained and fine-tuned EfficientNet and SeResNeXt architectures, with resolutions ranging from 224×224 to 450×450 pixels on the ISIC 2016–2018 datasets. They used cropping and resizing strategies to resize the images to the corresponding sizes. For non-square images, the remaining image space was zero-padded for the resizing strategy, and for the cropping strategy, the images were randomly cropped. The results showed that the cropping strategy performed better at all resolutions. Additionally, EfficientNetB0 achieved the highest Balanced Multi-Class Accuracy (BMCA) at 240×240 pixels, while EfficientNetB1 achieved the highest BMCA at 224×224 pixels, and SeResNeXt-50 achieved the highest BMCA at 380×380 pixels. Furthermore, they proposed a Multiscale Multi-CNN (MSM-CNN) fusion approach based on a three-level model ensemble strategy, which combined EfficientNetB0, EfficientNetB1, and SeResNeXt-50 architectures trained on cropped images at resolutions from 224×224 to 450×450 pixels. Their proposed MSM-CNN fusion approach outperformed all single models at all resolutions. The authors confirmed that image resolution affects skin lesion classification performance. Although outside the scope of their study, they conducted additional experiments on smaller resolutions of 32×32, 64×64, and 128×128 pixels, with results showing that decreasing resolution correlates with worse performance in terms of BMCA, indicating information loss at those smaller scales. In another study, Mahbod et al. [39] investigated the classification performance of fine-tuned and pre-trained ResNet and DenseNet architectures on different image resolutions of skin lesions, including malignant melanomas and seborrheic keratosis, ranging from 64×64 to 768×768 pixels on the ISIC 2016–2017 datasets. The results showed that models trained on images with resolutions of 64×64 pixels performed significantly worse than those trained on resolutions of 128×128 pixels and above. Furthermore, larger images correlated with better AUROC scores, although the performance increase was only slight. Consequently, the researchers concluded that an

increase in image resolution is correlated with enhanced classification outcomes, whereas downscaling results in the loss of information, as demonstrated in another study [38].

2.7 Research Question

Building on the findings of Mahbod et al. [38, 39], it is evident that image resolution and aspect ratio are critical factors in medical image classification. However, in their study [39], Mahbod et al. did not maintain the aspect ratio of non-square images, which could theoretically result in a loss of medically relevant details due to distortions by downscaling, potentially influencing the classification rate of the deep learning models. In another study, Mahbod et al. [38] applied zero padding to non-square images before downscaling them, thereby avoiding distortions. In addition to the zero padding resizing strategy, they also used a random cropping strategy. However, after observing that the cropping strategy yielded better results in early experiments, they opted to use it for further model training. This approach, while initially promising, can lead to an unintended emphasis on non-lesion details or the cropping of relevant skin lesion information, ultimately a□ecting classification performance and potentially leading to incorrect classification of skin lesions.

While in their study [38], Mahbod et al. investigated the e□ects of resolutions from 224×224 to 450×450 pixels, taking the aspect ratio into account, they suggested that further research on the e□ect of image sizes on larger scales can be conducted. Similarly, in another study [39], Mahbod et al. examined resolutions ranging from 64×64 to 768×768 pixels, however without maintaining the aspect ratio. This leaves a gap in understanding the e□ects of maintaining the aspect ratio for resolutions above 450×450 pixels.

Therefore, this study seeks to address this gap by investigating the following research question:

> How does maintaining the aspect ratio in high-resolution images (above 450×450 pixels) a□ect the classification performance of deep learning models in the detection of skin lesions?

To examine the research question, experiments are conducted by fine-tuning pre-trained ImageNet deep learning models based on the ResNet architecture [26] on three di□erent image resolutions: 300×225, 450×338, and 600×450 pixels. These models are trained

on a derived version of the ISIC 2019 dataset. Prior to investigating the effects of different image resolutions, experiments are conducted to determine the most effective augmentation method for training residual neural networks on the dataset, providing a more robust foundation for understanding the effect of image resolution.

3 Methodology

This chapter outlines the methodologies employed in preparing the dataset, pre-processing the images, creating the subsets for the experiments, data augmentation methods and techniques, and designing the model architecture. Furthermore, the chapter outlines the methodology employed in conducting the experiments, which are designed to determine the most e□ective augmentation method and address the research question regarding the e□ect of di□erent image resolutions on the classification rate of skin lesion models. The ISIC 2019 challenge dataset is utilized and revised for this purpose. Throughout this study, these experiments will be referred to as *E1* for the data augmentation experiments and *E2* for the image resolution experiments.

Section 3.1 presents the derived dataset utilized in this study, its class distribution, and the division into training, validation, and test sets. Furthermore, Section 3.2 outlines the various pre-processing steps necessary to reduce the existing artifacts in the dataset to improve the performance of the models in the experiments. The data augmentation pipeline applied to the training sets is presented in Section 3.3. Section 3.4 presents the model architecture, which is trained in all experiments using transfer learning and fine-tuning techniques. Finally, Section 3.5 provides a detailed description of the experimental procedures and the specific training procedures and hyperparameters for the experiments *E1* and *E2*.

3.1 Dataset Summary

In Section 3.1.1, the revised dataset used in this study, along with the class distribution, is discussed. Section 3.1.2 provides further details on the splitting of the dataset into training, validation, and test sets for the experiments.

3.1.1 Derived Dataset

As previously stated in Section 2.1, the ISIC 2019 dataset exhibits a high imbalance in the number of instances per class. To eliminate any potential bias towards the majority class and to ensure a more balanced experimental setup, the dataset is revised to ensure a more balanced distribution of classes. Specifically, the classes AK, SCC, VASC, and DF are being removed due to their significantly lower representation compared to the classes MEL, NV, BCC, and BKL (see Table 3.1). Additionally, the class NV is randomly downsampled to half its original size, reducing it from 12,875 to 6,437 samples, to achieve a more balanced representation among the classes, while no resampling is applied to the remaining classes. The derived dataset used in this study comprises a total of 16,906 images, as shown in Table 3.1.

Class	NV	MEL	BCC	BKL	AK	SCC	VASC	DF	**Total**
ISIC Dataset	12,875	4,522	3,323	2,624	867	628	253	239	**25,331**
Ratio %	50.83	17.85	13.12	10.36	3.42	2.48	1.00	0.94	**100%**
Derived Dataset	6,437	4,522	3,323	2,624	-	-	-	-	**16,906**
Ratio %	38.08	26.75	19.66	15.52	-	-	-	-	**100%**

Table 3.1: Class distribution and ratio of the ISIC 2019 dataset and the subset used in this study.

3.1.2 Dataset Splitting

At the time of writing, the ISIC has only released the test set for the 2019 challenge, lacking corresponding ground truth data. Additionally, no validation set is currently available. Consequently, the derived dataset (see Table 3.1) is split into training, validation, and test sets using a stratified train–validation–test split (70%/15%/15%) (see Table 3.2). Stratified splitting ensures that each target class is represented in proportions similar to those of the complete dataset. Both experiments *E1* and *E2* will utilize the stratified sets shown in Table 3.2 as a baseline for training.

Class	NV	MEL	BCC	BKL	**Total**
Training Set	4,506	3,165	2,326	1,837	**11,834**
Ratio %	38.08	26.74	19.66	15.52	**100%**
Validation Set	966	678	498	394	**2,536**
Ratio %	38.09	26.74	19.64	15.54	**100%**
Test Set	965	679	499	393	**2,536**
Ratio %	38.05	26.77	19.68	15.50	**100%**

Table 3.2: Class distribution and ratio of the stratified training, validation, and test sets of the derived dataset.

3.2 Image Pre-processing

Images with various artifacts, including DCAs (see Figure 3.1a), clinical markings (see Figure 3.1b), circular size reference stickers (see Figure 3.1c), and physical rulers (see Figure 3.1d), are present in the ISIC 2019 dataset. These artifacts do not provide relevant information for skin lesion classification, especially DCAs, which can negatively impact the performance of CNNs [63]. Therefore, it is necessary to pre-process these images to minimize the e□ects of these artifacts during the training, validation, and testing of CNNs. The pre-processing of the DCAs in this study follows a similar approach to that described by Gessert et al. [22].

(a) (b) (c) (d)

Figure 3.1: Various image artifacts present in the ISIC 2019 dataset: (a) Presence of a DCA, (b) presence of a clinical marking, (c) presence of a circular size reference sticker, and (d) presence of a ruler.

The phenomenon of DCAs is a consequence of vignetting, a characteristic that occurs when using tubular lenses in dermatoscopes [51]. This results in images exhibiting a relatively large proportion of black pixels, which must be removed to retain only the skin

lesion. Accordingly, any images with a proportion of black pixels in the decimal RGB value range of *0* to *30*, exceeding 15% of the total image area, will be automatically identified and marked for further pre-processing. In total, 2,323 images were identified as having DCAs. Following a manual review, 12 images were incorrectly identified, and 48 additional images were found that could not be identified by the black portion approach because they had only a small DCA or the intensity of the DCA was outside the threshold.

To remove most of the DCA area, it is necessary to identify the minimum enclosing rectangle around the circular skin lesions. For this purpose, a binary mask (see Figure 3.4c) is calculated for each image to obtain the circular skin lesion area. The images are first converted to grayscale, and a global threshold of *25* is applied to each pixel. If the pixel intensity falls below the threshold, it is set to *0* (representing black); otherwise, it is set to *1* (representing white). Subsequently, the resulting binary images (see Figure 3.4b) are morphologically modified. Morphological transformations modify the structure of an input image through the application of a structuring element (SE), which is defined as an $n \times m$ matrix, where n and m are integers. The SE superimposes the pixels over the input image, moving from left to right and top to bottom. The value of the central pixel superimposed by the SE is updated to either *0* or *1* based on the SE and the neighboring pixel values.

Following the application of the thresholding technique, the region enclosed by the circle may exhibit the presence of black pixel clusters or noise (see Figure 3.4b). These elements must be eliminated in order to facilitate the subsequent identification of the minimum enclosing rectangle surrounding the white circle. To remove these areas, a dilation operation (see Figure 3.2) is performed using a 27×27 SE to expand the white regions. Dilation [71] is a morphological operation defined as

$$A \oplus B = \bigcup_{b \in B} A_b, \tag{3.1}$$

that expands the boundaries of the foreground object in an image by setting the central pixel to *1* if at least one pixel under the SE is *1*. However, small black noises may remain after dilation. To address this, a closing operation, which is dilation followed by erosion with a 9×9 SE, is applied to remove black noise within the circular area. Closing [71] is defined as

$$A \bullet B = (A \oplus B) \ominus B. \tag{3.2}$$

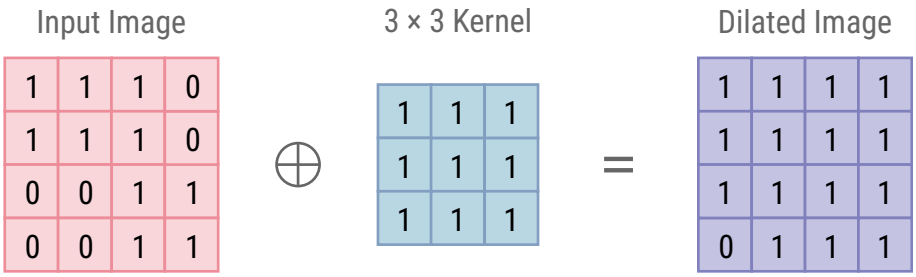


Figure 3.2: Effect of dilation using a 3×3 square structuring element.

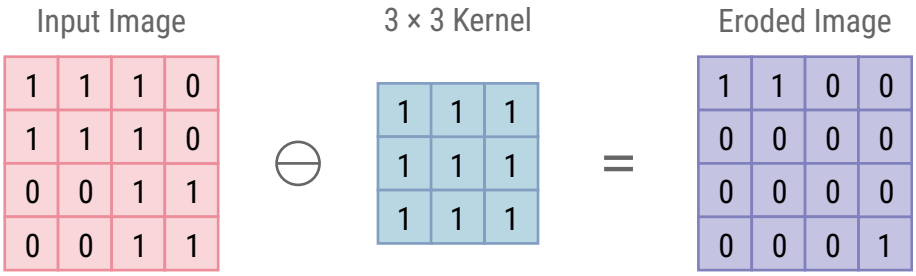


Figure 3.3: Effect of erosion using a 3×3 square structuring element.

Erosion [71] (see Figure 3.3) shrinks the boundaries of the foreground object and sets the central pixel to *0* if at least one pixel under the SE is *0*. Erosion is defined as

$$A \ominus B = \bigcap_{b \in B} A_{-b}. \tag{3.3}$$

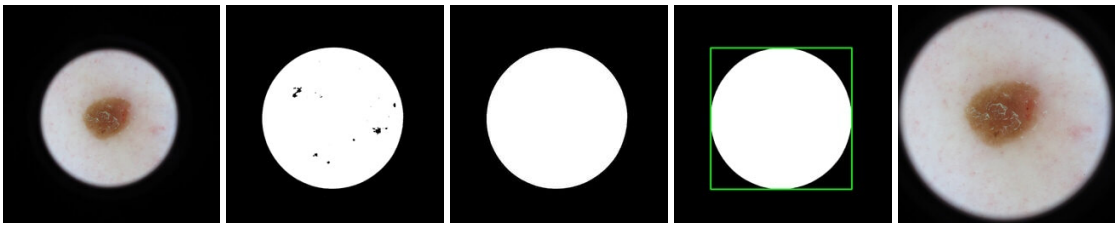
Subsequently, an opening operation, which is erosion followed by dilation using a 12×12 SE, is performed to remove white noise outside the circular area. Opening [71] is defined as

$$A \circ B = (A \ominus B) \oplus B. \tag{3.4}$$

The binary mask (see Figure 3.4c) is then used to determine the minimum enclosing rectangle around the boundaries of the circular shape. The circular shape can be located by applying the Ramer–Douglas–Peucker contour approximation algorithm from the Open Source Computer Vision Library (OpenCV) library[1]. The outermost contours

[1] https://opencv.org/ (Accessed 05/18/24)

obtained from this contour approximation are then used to calculate the minimum enclosing rectangle encompassing the circular shape (see Figure 3.4d). The coordinates of this rectangle are subsequently applied to the DCA to extract the skin lesion, as shown in Figure 3.4e.



(a) (b) (c) (d) (e)

Figure 3.4: Pre-processing steps to remove DCAs from images in the derived dataset. (a) Original image, (b) binary mask after thresholding, (c) resulting binary mask after thresholding and morphological operations, (d) minimum enclosing rectangle around the circular shape, and (e) extraction of the skin lesion based on the minimum enclosing rectangle coordinates of (d) applied to (a).

The identification and automatic removal of artifacts from Figures 3.1b to 3.1d presents a more significant challenge due to the unpredictable locations of these artifacts. An interesting feature of the ISIC 2019 dataset is that skin lesions are consistently located at the center of the images, whereas artifacts tend to appear more frequently at the boundaries. To minimize the potential interference of boundary artifacts, all images in the training set were automatically center-cropped to 600×450 pixels, which represents the highest common resolution among the classes used in the derived dataset, as stated in Section 2.1. Upon manual review, 382 images were found to contain skin lesions extending beyond the cropped area. These images were subsequently manually cropped and corrected to ensure the lesions were properly centered. After removing the artifacts and cropping, three training subsets were created for the experiments *E1* and *E2*, each with di□erent resolutions. The original 600×450 pixel training images were downscaled to 450×338 and 300×225 pixels, respectively, while preserving the original aspect ratio. The downscaling process was performed using the implementation of the Lánczos algorithm from the Pillow library[2].

[2] https://python-pillow.org/ (Accessed 05/18/24)

3.3 Data Augmentation

Data augmentation is used to increase the size and diversity of the data while reducing overfitting and improving generalization performance [61]. In the experiments, both offline and online data augmentation methods are employed. Offline augmentation involves augmenting the images and adding them to the dataset prior to training. Conversely, online augmentation is applied in real-time before each epoch during training.

(a) (b) (c) (d) (e) (f) (g)

Figure 3.5: Data augmentation techniques applied to the training sets. (a) Without augmentation, (b) rotation transformation ($+60°$), (c) flipping transformation (horizontal), (d) shearing transformation ($+10°$), (e) shifting transformations ($+10\%$ width and height), (f) zooming transformation ($+20\%$), and (g) brightness and contrast adjustments (-20% brightness, $+20\%$ contrast).

The augmentation pipeline incorporates several techniques to enhance the dataset. Initially, images are without any augmentation (see Figure 3.5a). These images then undergo random rotations within a range of $\pm 180°$ (see Figure 3.5b). Additionally, they are flipped horizontally and vertically with a probability of 50% (see Figure 3.5c). Shearing along the x-axis is applied with an intensity of up to $0.2°$ (see Figure 3.5d). Images are also shifted in width and height within a range of $\pm$10% (see Figure 3.5e), and zooming in is performed up to 20% (see Figure 3.5f). Furthermore, brightness and contrast are adjusted within a range of $\pm$20% (see Figure 3.5g).

Throughout these affine transformations, nearest-neighbor interpolation is used to fill any empty areas with the nearest pixel values. All input images from the training, validation, and test sets are rescaled to a range of $[0, 1]$. Since the models used in the experiments are pre-trained on the ImageNet dataset, channel-wise normalization is performed using the ImageNet mean and standard deviation values: mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$. This normalization helps improve training convergence and facilitates more effective learning by aligning the features of the input images with those used in the ImageNet dataset.

3.4 Model Architecture

The models utilized in the experiments comprise two main components: the body and the head (see Figure 3.6). The body is responsible for the extraction of features from the input images and comprises convolutional and pooling layers. Convolutional layers apply filters to the images to detect features such as edges, textures or shapes. Pooling layers then reduce spatial dimensions while preserving important features. In conclusion, the head is responsible for generating the final result based on the extracted features of the body. This is achieved through the use of fully connected layers, which are followed by an output layer corresponding to the number of classes.

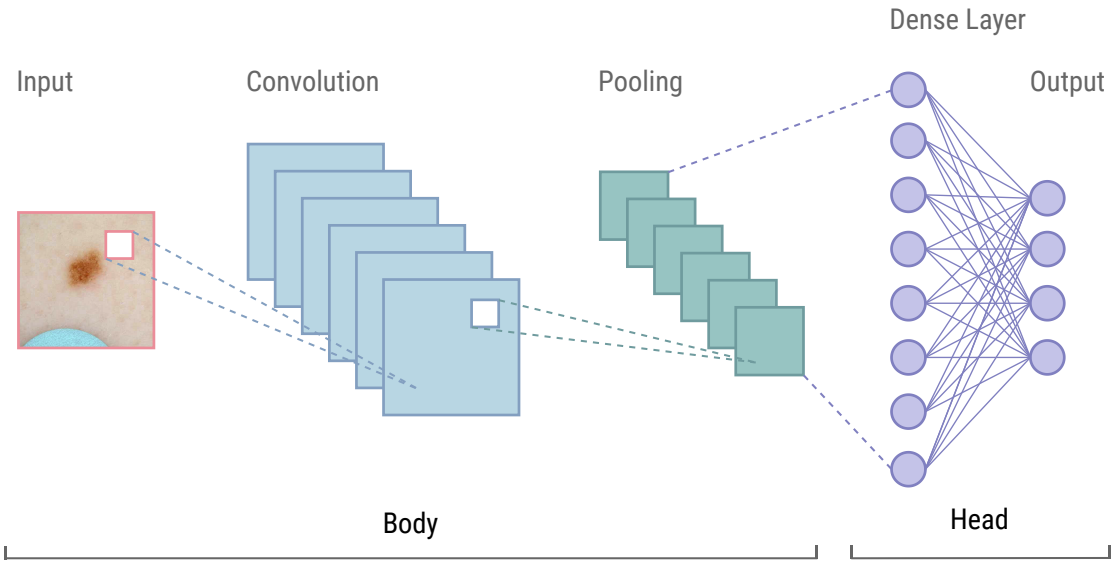


Figure 3.6: The standard CNN architecture comprises two main components: the body, which comprises convolutional layers for feature extraction and pooling layers for spatial reduction, and the head, which consists of dense layers for classification and the output layer for generating predictions (adapted from Phung and Rhee [52] p. 2).

The ResNet-50 architecture [26], which had been pre-trained on ImageNet, was chosen as the base architecture for all experiments. Since the model weights were trained on images with a resolution of 224×224 pixels and do not accommodate other input sizes, only the ResNet-50 body is used as a feature extractor, while the ResNet-50 head is replaced by a modified variant to allow transfer learning on different image sizes.

The modified head comprises a Global Average Pooling layer, which computes the average value of each feature map across its spatial dimensions. Followed by a Dense layer comprising 1,024 neurons, which are activated by Rectified Linear Unit (ReLU). Additionally, L2 regularization with a regularization factor of 1.0×10^{-2} is used to prevent overfitting [76] by penalizing large weights. The Batch Normalization layer then normalizes the activations of the previous layer. Furthermore, a Dropout layer with a probability of $p = 0.5$ is then applied to prevent overfitting [65]. The final layer is a Dense layer comprising four neurons and activated by the Softmax activation function. This function transforms the output values into a probability distribution across four classes, ensuring that the sum of the probabilities within the interval of $[0, 1]$ is equal to *1*. A detailed overview of the complete modified head variant and body architecture is presented in Table 3.3.

3.5 Training Procedure

The training procedures are divided into three subsections. The first subsection 3.5.1 outlines the general training procedure, including the hyperparameters used, which are applied to both experiments. The second subsection 3.5.2 describes the specific procedure for *E1*, which aims to determine the most e□ective augmentation method on the dataset employed. The final subsection 3.5.3 presents the procedure for *E2*, which addresses the central research question.

3.5.1 General Training Parameters

The model architecture outlined in Section 3.4 employs a batch size of 32. Initially, transfer learning is utilized by freezing the body layers while the head is initialized with He–Initialization [25] and trained for five epochs. Subsequently, the body layers are unfrozen and fine-tuned for up to 60 additional epochs. If the AUROC score and loss do not improve for more than 15 consecutive epochs, early stopping is triggered. The model with the highest AUROC score and the lowest loss achieved throughout the entire training period is then saved. The selection of hyperparameters and their respective values follows established standards and is influenced by prior studies [13, 37, 38, 60]. As a result, the Adam Optimizer was chosen with an initial learning rate of 1.0×10^{-4}, a Beta 1 value of $\beta_1 = 0.90$, and a Beta 2 value of $\beta_2 = 0.99$. Furthermore, the learning

Part	Type	Kernel Size	Filter Number	Output Size *Model A*	Output Size *Model B*	Output Size *Model C*	Times
BODY	Conv 1	7×7	64	$300 \times 225 \times 64$	$225 \times 169 \times 64$	$150 \times 113 \times 64$	1
	Max Pooling	3×3	-	$150 \times 113 \times 64$	$113 \times 85 \times 64$	$75 \times 57 \times 64$	1
	Conv 2	1×1	64	$150 \times 113 \times 64$	$113 \times 85 \times 64$	$75 \times 57 \times 64$	3
		3×3	64	$150 \times 113 \times 64$	$113 \times 85 \times 64$	$75 \times 57 \times 64$	
		1×1	256	$150 \times 113 \times 256$	$113 \times 85 \times 256$	$75 \times 57 \times 256$	
	Conv 3	1×1	128	$75 \times 57 \times 128$	$57 \times 43 \times 128$	$38 \times 29 \times 128$	4
		3×3	128	$75 \times 57 \times 128$	$57 \times 43 \times 128$	$38 \times 29 \times 128$	
		1×1	512	$75 \times 57 \times 512$	$57 \times 43 \times 512$	$38 \times 29 \times 512$	
	Conv 4	1×1	256	$38 \times 29 \times 256$	$29 \times 22 \times 256$	$19 \times 15 \times 256$	6
		3×3	256	$38 \times 29 \times 256$	$29 \times 22 \times 256$	$19 \times 15 \times 256$	
		1×1	1024	$38 \times 29 \times 1024$	$29 \times 22 \times 1024$	$19 \times 15 \times 1024$	
	Conv 5	1×1	512	$19 \times 15 \times 512$	$15 \times 11 \times 512$	$10 \times 8 \times 512$	3
		3×3	512	$19 \times 15 \times 512$	$15 \times 11 \times 512$	$10 \times 8 \times 512$	
		1×1	2048	$19 \times 15 \times 2048$	$15 \times 11 \times 2048$	$10 \times 8 \times 2048$	
HEAD	Global Average Pooling	-	-	2048	2048	2048	1
	Dense 1 ReLU, L2	-	1	1024	1024	1024	1
	Batch Norm	-	1	1024	1024	1024	1
	Dropout	-	1	1024	1024	1024	1
	Dense 2 Softmax	-	1	4	4	4	1

Table 3.3: The ResNet-50 architectures used in the experiments, including the corresponding output sizes after each layer (height $\times$ width $\times$ channels). *Model A* with an input shape of $(600 \times 450 \times 3)$, *Model B* with an input shape of $(450 \times 338 \times 3)$, and *Model C* with an input shape of $(300 \times 225 \times 3)$.

rate is reduced dynamically when no improvement is observed in either the AUROC score or loss for a period of over five epochs by a reduction factor of 2.0×10^{-1} until reaching a minimum learning rate of 1.0×10^{-6}.

Depending on the augmentation method applied, distinct loss functions are employed. For o□ine augmentation, the categorical cross-entropy loss function is utilized, while for online augmentation, the categorical focal loss function proposed by Lin et al. [34] is employed. Categorical cross-entropy calculates the loss based on the di□erence between the predicted probability distribution and the true distribution for each sample, but it does not consider the imbalance in the class distribution. The categorical cross-entropy loss is defined as

$$\mathsf{CEL}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{3.5}$$

where y_i is the true label of the class i, $\hat{y}_i$ is the predicted probability of the class i, and N is the number of classes.

Categorical focal loss addresses the imbalance in the training set by down-weighting the loss contribution from the samples that are easy to classify and increasing the weight for the hard samples. This focuses more on samples that are di□cult to classify, which helps in learning from minority classes and improving overall model performance. Categorical focal loss is defined as

$$\mathsf{FL}(y, \hat{y}) = -\sum_{i=1}^{N} \alpha_i (1 - \hat{y}_i)^{\gamma} y_i \log(\hat{y}_i), \tag{3.6}$$

where α_i is the weighting factor, which is defined as the inverse of the frequency of class i, and $(1 - \hat{y}_i)^{\gamma}$ the modulating factor with the focusing parameter gamma $\gamma = 2$. The focusing parameter was selected based on the results obtained in the experiments reported by Lin et al. [34], which yielded the best results.

Across all experiments, to maintain consistency, the same architectures shown in Table 3.3 are employed, with consistent hyperparameters, augmentation pipeline, and seeds used for all sub-experiments to introduce controlled randomness and variability. This approach ensures that changes in the AUROC scores in *E1* are dependent on the augmentation method, while for *E2* the changes are attributable to the variation in image resolutions, with all evaluations performed on the test set.

The experiments are all conducted using NVIDIA Quadro P6000 $24\,\mathrm{GB}$ GPUs, provided by the Creative Space for Technical Innovations (CSTI)[3] as part of this study. Additionally, the models are implemented using the deep learning API Keras[4] 3.0 on top of the TensorFlow[5] framework version 2.15 and CUDA[6] 12.2.

3.5.2 Data Augmentation Experiment

E1 is comprised of two sub-experiments, with a total of 20 trained models. Each sub-experiment employs the training subset with an image resolution of 600×450 pixels and the pre-trained *Model A* (see Table 3.3). In the first sub-experiment, ten models are trained on ten distinct o□ine augmented training sets. Each training set is balanced to maintain a class distribution of 4,506 images per class, which corresponds to the number of images in the majority class NV. The second sub-experiment utilizes online augmented training sets with the same distinct seeds during the training of the ten models.

3.5.3 Image Resolution Experiment

E2 is comprised of three sub-experiments, with a total of 75 trained models. The method of augmentation to be employed for training the models will depend on the outcome of *E1*. This will either be o□ine augmentation or online augmentation. In both scenarios, the pre-trained models *Model A*, *Model B*, and *Model C* (see Table 3.3) will be employed. Each training subset, corresponding to the input shape of models with resolutions of 600×450, 450×338, and 300×225 pixels, will be augmented with the same 25 distinct seeds, resulting in 25 trained models in total for each resolution. For the o□ine augmentation, the sets will be balanced to the same distribution of 4,506 images per class as in *E1*.

[3] https://csti.haw-hamburg.de/ (Accessed 05/18/24)
[4] https://keras.io/ (Accessed 05/19/24)
[5] https://tensorflow.org/ (Accessed 05/22/24)
[6] https://developer.nvidia.com/cuda-toolkit (Accessed 05/22/24)

4 Evaluation

This chapter presents an evaluation of the two experiments carried out. The performance of the individual models is evaluated using the AUROC score, which is calculated using the ISIC challenge scoring repository[1].

In both experiments, models achieving the lowest training loss and highest AUROC score are selected. For empirical evaluation, all models are evaluated on the test dataset from Section 3.1.2. The mean and one standard deviation of AUROC scores are calculated for each model group, and results are visualized in graphs.

Initially, in Section 4.1, the results of the experiments *E1* and *E2* are presented. Finally, Section 4.2 discusses the results obtained from the experiments. First, the relationship between the AUROC score and the augmentation methods employed in *E1* is analyzed to determine the most suitable method for the selected dataset. Furthermore, the effects of image resolution on the classification rate of the AUROC score are examined. Here, the models are augmented using the method that proved more effective in the preceding experiment.

4.1 Results

The following Section 4.1.1 presents the results of the experiment *E1* regarding data augmentation. The results of the image resolution experiment *E2* are presented in Section 4.1.2.

[1] https://github.com/ImageMarkup/isic-challenge-scoring/tree/master (Accessed 06/01/24)

4.1.1 Data Augmentation Experiments

This section presents the results of *E1*. In both sub-experiments, a total of 20 distinct models were trained on the same 10 seeds, each based on the model architecture of *Model C*. The graphs in Figures 4.1a to 4.1d show that at a resolution of 600×450 pixels, online augmentation achieved higher AUROC scores compared to o□ine augmentation across all classes. In the NV and BCC classes, the mean di□erence Δ between o□ine and online augmentation was $\Delta\text{NV} = 0.0167$ and $\Delta\text{BCC} = 0.0144$, respectively, while in the BKL and MEL classes, the di□erence was notably larger at $\Delta\text{BKL} = 0.0281$ and $\Delta\text{MEL} = 0.0409$. Furthermore, the variability of AUROC scores, as indicated by the standard deviation, was smaller for all classes using online augmentation compared to o□ine augmentation. Notably, in the NV, BCC, and BKL classes, the variability of the AUROC scores was clearly reduced when employing online augmentation. In the MEL class, despite the largest mean di□erence, the variability also decreased compared to o□ine augmentation.

4.1.2 Image Resolution Experiments

This section presents the results of *E2*, where a total of 75 models were trained using online augmentation at resolutions of 600×450, 450×338, and 300×225 pixels across three sub-experiments. These models were based on the architectures of *Models A* to *C* (see Table 3.3) and utilized the same 25 seeds for consistency. Graphs in Figures 4.2a to 4.2d illustrate a positive correlation between resolution and AUROC scores. Across all classes, the highest resolution 600×450 pixels yielded the highest mean AUROC scores, while the lowest resolution 300×225 pixels resulted in the lowest mean scores. The intermediate resolution of 450×338 pixels produced mean scores that fell between these two extremes. The smallest di□erences in mean AUROC scores between the lowest resolution of 300×225 pixels and the highest resolution of 600×450 pixels in one class were observed in the BCC and NV classes, with $\Delta\text{BCC} = 0.0091$ and $\Delta\text{NV} = 0.0124$. Conversely, the MEL and BKL classes exhibited larger di□erences, with $\Delta\text{MEL} = 0.0202$ and $\Delta\text{BKL} = 0.0285$. Additionally, the variability in AUROC scores, as measured by one standard deviation, was small and decreased with higher resolution in the BCC and NV classes. In contrast, the classes MEL and BKL, showed a more pronounced initial spread that also decreased with increasing resolution.

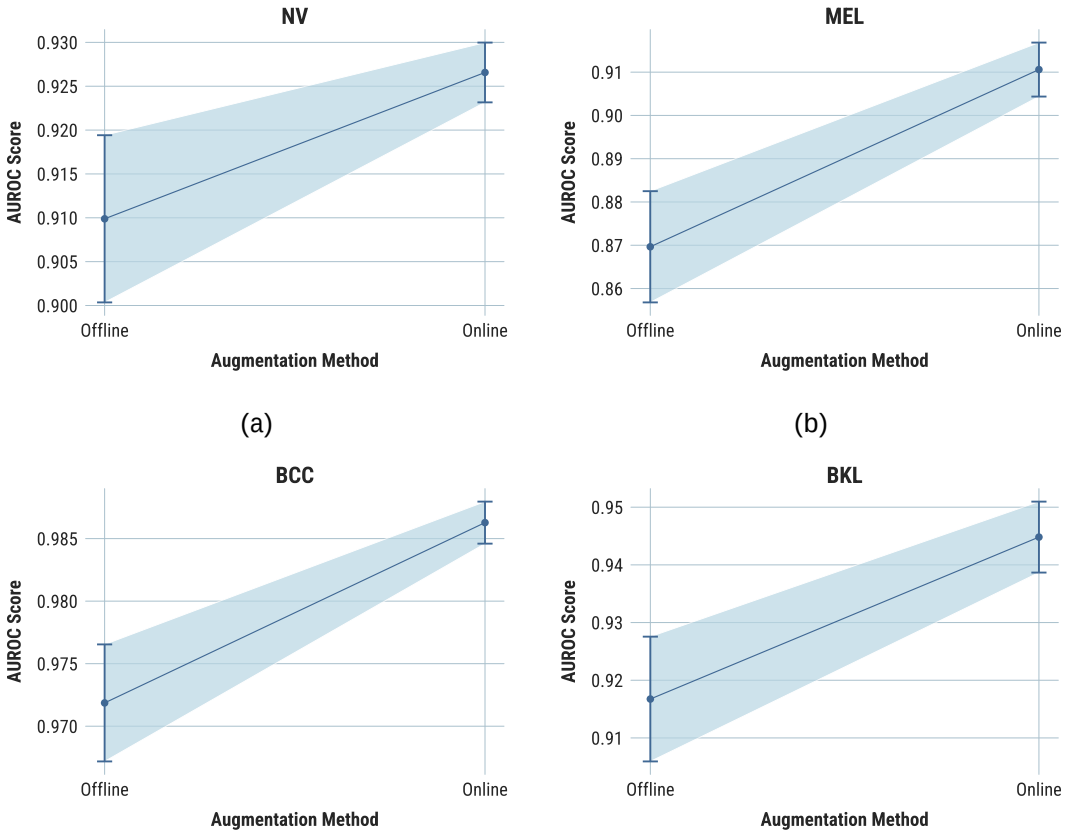


Figure 4.1: Mean AUROC scores and their standard deviations (± 1 SD) at 600×450 pixels, obtained using both o□ine and online augmentation methods across 10 models each. (a) Class NV, (b) class MEL, (c) class BCC, and (d) class BKL.

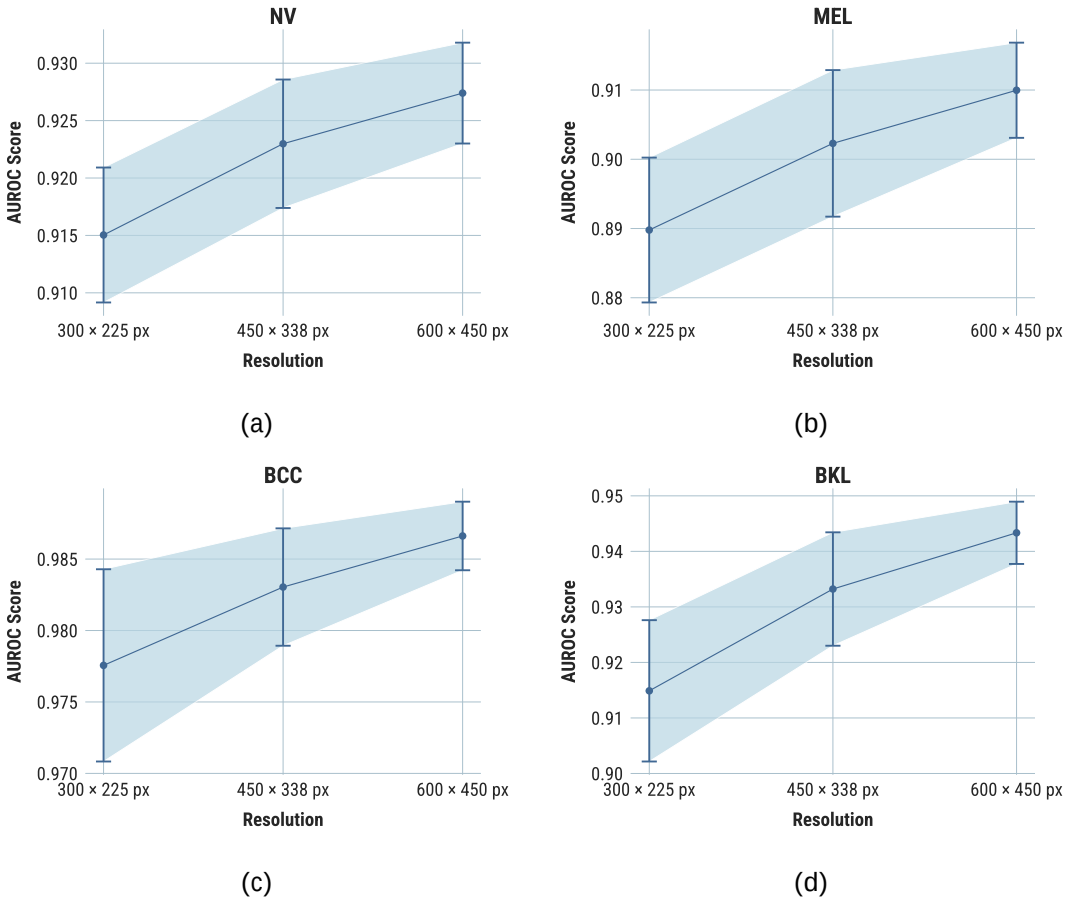


Figure 4.2: Mean AUROC scores and their standard deviations (± 1 SD) at 300×225, 450×338, and 600×450 pixels, across 25 models per resolution. (a) Class NV, (b) class MEL, (c) class BCC, and (d) class BKL.

4.2 Discussion

E1 evaluated the effects of online versus offline data augmentation on the performance of models trained on the *Model C* (see Table 3.3) architecture at a resolution of 600×450 pixels. A total of 20 distinct models were trained across two sub-experiments, with 10 models in each sub-experiment. In both sub-experiments, the same set of 10 different seeds was used, ensuring the robustness of the results through repeated trials. The performance metric utilized was the AUROC metric, evaluated across four classes of skin lesions: NV, BCC, BKL, and MEL.

The results of *E1*, presented in Figures 4.1a to 4.1d, clearly demonstrate the superior performance of online data augmentation compared to offline augmentation. Online augmentation consistently yielded higher mean AUROC scores across all classes. For instance, the mean differences between offline and online augmentation were $\Delta\text{NV} = 0.0167$ and $\Delta\text{BCC} = 0.0144$ for the NV and BCC classes, respectively. The BKL and MEL classes exhibited even more substantial improvements, with mean differences of $\Delta\text{BKL} = 0.0281$ and $\Delta\text{MEL} = 0.0409$, respectively. This indicates a particularly strong effect of online augmentation on enhancing performance in these classes. Furthermore, the analysis of the variability provided significant insights. The variability was significantly lower for all classes with online augmentation compared to offline augmentation. The NV, BCC, and BKL classes all demonstrated reduced variability with online augmentation. Notably, even the MEL class, despite exhibiting a larger mean improvement, experienced less variability with online augmentation compared to offline augmentation. These findings demonstrate the effectiveness of online augmentation in enhancing the performance and robustness of classification models by reducing variability across different classes. The greater variations in the augmented data provided by online augmentation likely facilitate model generalization, enabling the capture of nuanced inter-class similarities and intra-class variation features of skin lesions, thereby improving classification accuracy and consistency.

E2 investigated the effect of image resolution at 600×450, 450×338, and 300×225 pixels on model performance in the classification of skin lesions. Based on the results of *E1*, where online augmentation achieved better results than offline augmentation, online augmentation was used in this experiment. A total of 75 models, based on the architectures of *Models A* to *C* (see Table 3.3), were trained at these three different resolutions. The same 25 seeds were used for each resolution experiment to ensure consistency.

The results, presented in Figures 4.2a to 4.2d, reveal a clear and strong positive correlation between image resolution and AUROC scores. The highest resolution (600×450 pixels) consistently produced the highest mean AUROC scores across all classes of skin lesions. In contrast, the lowest resolution (300×225 pixels) resulted in the lowest mean scores, with the intermediate resolution (450×338 pixels) yielding mean scores between these two extremes. The magnitude of the difference in mean AUROC scores between the highest and lowest resolutions varied by class. The smallest differences were observed in the BCC and NV classes, with $\Delta\text{BCC} = 0.0091$ and $\Delta\text{NV} = 0.0124$. Conversely, the MEL and BKL classes exhibited larger differences, with $\Delta\text{MEL} = 0.0202$ and $\Delta\text{BKL} = 0.0285$, further emphasizing the resolution-dependence of model performance. Moreover, the variability in AUROC scores decreased with increasing resolution for all classes. The BCC and NV classes showed relatively small initial variability, which further diminished at higher resolutions. In contrast, the MEL and BKL classes had larger initial spreads in AUROC scores, but these also reduced as resolution increased.

From a medical perspective, these results highlight the significance of using high-resolution images in achieving both higher performance and reduced variability in the automated diagnosis of skin lesions. This is particularly important for MEL and BKL, where higher-resolution images significantly improve diagnostic performance, potentially leading to better patient outcomes through early and accurate detection. The observed decline in performance in relation to image resolution is likely to be a consequence of information loss resulting from downscaling. This process makes it more challenging for the models to distinguish between various skin lesions as the inter-class similarity and intra-class variation similarity features become more pronounced. Another factor that can influence the results is the model architecture and depth, as well as the choice of hyperparameters and the loss function used. These factors can affect the ability of deep learning models to differentiate between classes and handle variations in image quality and resolution.

One limitation of this study is that the effects of image resolution were only examined for images with a maximum resolution of 600×450 pixels. This was due to the fact that the dataset comprised images of varying resolutions, with a resolution of 600×450 pixels being the highest common resolution among the classes used. Additionally, the dataset exhibited a significant class imbalance. As a result, classes with fewer than 1,000 images were excluded to achieve a more balanced dataset and to prevent the models from being biased towards the majority classes. Another limitation is the use of pre-trained models based on ImageNet. As these models were trained on natural images and did not include domain-specific medical features, this may have affected their performance and

generalizability. Furthermore, the exploration of additional advanced data augmentation techniques, such as GANs or alternative pre-processing techniques, could potentially enhance the performance and robustness of the models. Moreover, this study employed hyperparameters that adhered to established standards. However, optimizing hyperparameters could potentially result in a significant performance improvement. Techniques such as grid search, random search, and Bayesian optimization o□er e□ective methods for fine-tuning parameters, including learning rate, batch size, and layer configurations. A systematic exploration of these parameters may identify optimal settings, potentially resulting in substantial performance improvements.

5 Summary & Outlook

This thesis investigated the e□ects of image resolution and data augmentation on the classification performance of skin lesions using deep learning models. The primary objective was to address the research gap identified by Mahbod et al. [38] regarding the optimal cropping or scaling factor for images at resolutions of 450×450 pixels and above, while maintaining the aspect ratio. Extensive experiments were conducted on a revised version of the ISIC 2019 dataset to identify the most e□ective data augmentation method, aiming to enhance the performance in subsequent experiments focused on image resolution. The ResNet-50 architecture, pre-trained on ImageNet, was utilized for all experiments.

The results indicated that online augmentation significantly outperforms o□ine augmentation in enhancing model performance across all classes at a resolution of 600×450 pixels. This method not only improved the generalization of the deep learning models but also enhanced their robustness by reducing the variability across all classes of skin lesions. Maintaining aspect ratios in skin lesion images at resolutions of 300×225, 450×338, and 600×450 pixels demonstrated a strong positive correlation with improved classification performance and reduced variability, as reflected in consistently higher and more closely clustered AUROC scores across all classes. This underscores the critical importance of detailed image information for accurate classification. Notably, the classes MEL and BKL exhibited the most significant improvements with higher resolutions, suggesting that these classes benefit the most from detailed image information. Conversely, lower resolutions exhibited decreased performance in all classes, aligning with previous findings by Mahbod et al. [38] on resolutions such as 32×32, 64×64, and 128×128 pixels, where diminished BMCA scores were observed due to information loss.

Future research could explore several avenues to enhance the classification accuracy of skin lesion images. First, investigating higher image resolutions exceeding 600×450 pixels while carefully maintaining the aspect ratio could provide more detailed information for the models. Additionally, addressing the limitations mentioned in this study, future experiments should include classes that were previously excluded due to a limited number

of images. This can be achieved through upsampling techniques, which may significantly affect performance by providing a more balanced dataset. Another promising area is the generation of synthetic images using GANs or diffusion models. It would be beneficial to assess how these synthetic images impact classification accuracy when mixed with existing data or used as replacements, both at the current and higher resolutions. Moreover, using models pre-trained on datasets specifically tailored to skin lesions instead of relying solely on ImageNet weights could enhance transfer learning outcomes. Models trained on skin lesion datasets are likely to learn more relevant and beneficial representations for dermatology-related tasks, potentially leading to better performance. Finally, it is crucial to examine the relationship between image resolution and classification accuracy relative to the computational cost of training these models. This consideration is particularly important for the future integration of these technologies into medical workflows, where both efficiency and ethical considerations related to patient health must be taken into account.

Bibliography

[1] Abhishek, Kumar ; Hamarneh, Ghassan: Mask2Lesion: Mask-Constrained Adversarial Skin Lesion Image Synthesis. In: Burgos, Ninon (Hrsg.) ; Gooya, Ali (Hrsg.) ; Svoboda, David (Hrsg.): *Simulation and Synthesis in Medical Imaging*. Cham : Springer International Publishing, 2019, S. 71–80. – URL https://doi.org/10.1007/978-3-030-32778-1_8. – ISBN 978-3-030-32778-1

[2] Afza, Farhat ; Sharif, Muhammad ; Khan, Muhammad A. ; Tariq, Usman ; Yong, Hwan-Seung ; Cha, Jaehyuk: Multiclass Skin Lesion Classification Using Hybrid Deep Features Selection and Extreme Learning Machine. In: *Sensors* 22 (2022), Nr. 3, S. 1–22. – URL https://doi.org/10.3390/s22030799. – ISSN 1424-8220

[3] Ahmad, Naveed ; Shah, Jamal H. ; Khan, Muhammad A. ; Baili, Jamel ; Ansari, Ghulam J. ; Tariq, Usman ; Kim, Ye J. ; Cha, Jae-Hyuk: A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. In: *Frontiers in Oncology* 13 (2023), S. 1–17. – URL https://doi.org/10.3389/fonc.2023.1151257. – ISSN 2234-943X

[4] Ali, Md S. ; Miah, Md S. ; Haque, Jahurul ; Rahman, Md M. ; Islam, Md K.: An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. In: *Machine Learning with Applications* 5 (2021), S. 1–8. – URL https://doi.org/10.1016/j.mlwa.2021.100036. – ISSN 2666-8270

[5] Alphonse, A. S. ; Benifa, J. V. B. ; Muaad, Abdullah Y. ; Chola, Channabasava ; Heyat, Md Belal B. ; Mursheed, Belal Abdullah H. ; Abdel Samee, Nagwan ; Alabdulhafith, Maali ; Al-antari, Mugahed A.: A Hybrid Stacked Restricted Boltzmann Machine with Sobel Directional Patterns for Melanoma Prediction in Colored Skin Images. In: *Diagnostics* 13 (2023), Nr. 6,

S. 1–24. – URL https://doi.org/10.3390/diagnostics13061104. – ISSN 2075-4418

[6] Alshahrani, Mohammed ; Al-Jabbar, Mohammed ; Senan, Ebrahim M. ; Ahmed, Ibrahim A. ; Mohammed Saif, Jamil A.: Analysis of dermoscopy images of multi-class for early detection of skin lesions by hybrid systems based on integrating features of CNN models. In: *PLOS ONE* 19 (2024), 3, Nr. 3, S. 1–36. – URL https://doi.org/10.1371/journal.pone.0298305

[7] Arnold, Melina ; Singh, Deependra ; Laversanne, Mathieu ; Vignat, Jerome ; Vaccarella, Salvatore ; Meheus, Filip ; Cust, Anne E. ; Vries, Esther de ; Whiteman, David C. ; Bray, Freddie: Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040. In: *JAMA Dermatology* 158 (2022), 05, Nr. 5, S. 495–503. – URL https://doi.org/10.1001/jamadermatol.2022.0160. – ISSN 2168-6068

[8] Borgli, Hanna ; Thambawita, Vajira ; Smedsrud, Pia H. ; Hicks, Steven ; Jha, Debesh ; Eskeland, Sigrun L. ; Randel, Kristin R. ; Pogorelov, Konstantin ; Lux, Mathias ; Nguyen, Duc Tien D. ; Johansen, Dag ; Griwodz, Carsten ; Stensland, Hakon K. ; Garcia-Ceja, Enrique ; Schmidt, Peter T. ; Hammer, Hugo L. ; Riegler, Michael A. ; Halvorsen, Pal ; Lange, Thomas de: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. In: *Scientific Data* 7 (2020), Nr. 283, S. 1–14. – URL https://doi.org/10.1038/s41597-020-00622-y. – ISSN 2052-4463

[9] Bozkurt, Ferhat: Skin lesion classification on dermatoscopic images using e□ective data augmentation and pre-trained deep learning approach. In: *Multimedia Tools and Applications* 82 (2023), Nr. 12, S. 18985–19003. – URL https://doi.org/10.1007/s11042-022-14095-1

[10] Carter, Jane V. ; Pan, Jianmin ; Rai, Shesh N. ; Galandiuk, Susan: ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. In: *Surgery* 159 (2016), Nr. 6, S. 1638–1645. – URL https://doi.org/10.1016/j.surg.2015.12.029. – ISSN 0039-6060

[11] Cassidy, Bill ; Kendrick, Connah ; Brodzicki, Andrzej ; Jaworek-Korjakowska, Joanna ; Yap, Moi H.: Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. In: *Medical Image analysis* 75 (2022),

S. 1–15. – URL https://doi.org/10.1016/j.media.2021.102305. – ISSN 1361-8415

[12] Chato, Lina ; Regentova, Emma: Survey of Transfer Learning Approaches in the Machine Learning of Digital Health Sensing Data. In: *Journal of Personalized Medicine* 13 (2023), Nr. 12, S. 1–53. – URL https://doi.org/10.3390/jpm13121703. – ISSN 2075-4426

[13] Chizari, Nima: *The Impact of Image Resolution and Model Scaling on Deep Learning based Automated Chest Radiograph Interpretation*. Hamburg, Germany, Hochschule für Angewandte Wissenschaften Hamburg, Master Thesis, November 2021. – URL https://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/master/chizari.pdf. – (Accessed 05/25/24)

[14] Codella, Noel C. F. ; Gutman, David ; Celebi, M. E. ; Helba, Brian ; Marchetti, Michael A. ; Dusza, Stephen W. ; Kalloo, Aadi ; Liopyris, Konstantinos ; Mishra, Nabin ; Kittler, Harald ; Halpern, Allan: Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, S. 168–172. – URL https://doi.org/10.1109/ISBI.2018.8363547. – ISSN 1945-8452

[15] Codella, Noel C. F. ; Rotemberg, Veronica ; Tschandl, Philipp ; Celebi, M. E. ; Dusza, Stephen W. ; Gutman, David A. ; Helba, Brian ; Kalloo, Aadi ; Liopyris, Konstantinos ; Marchetti, Michael A. ; Kittler, Harald ; Halpern, Allan: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). In: *CoRR* abs/1902.03368 (2019), S. 1–12. – URL https://doi.org/10.48550/arXiv.1902.03368

[16] Combalia, Marc ; Codella, Noel C. F. ; Rotemberg, Veronica ; Helba, Brian ; Vilaplana, Veronica ; Reiter, Ofer ; Carrera, Cristina ; Barreiro, Alicia ; Halpern, Allan C. ; Puig, Susana ; Malvehy, Josep: BCN20000: Dermoscopic Lesions in the Wild. (2019), S. 1–3. – URL https://doi.org/10.48550/arXiv.1908.02288

[17] Deng, Jia ; Dong, Wei ; Socher, Richard ; Li, Li-Jia ; Li, Kai ; Fei-Fei, Li: ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, S. 248–255. – URL

https://doi.org/10.1109/CVPR.2009.5206848. – ISBN 978-1-4244-3992-8

[18] Dildar, Mehwish ; Akram, Shumaila ; Irfan, Muhammad ; Khan, Hikmat U. ; Ramzan, Muhammad ; Mahmood, Abdur R. ; Alsaiari, Soliman A. ; Saeed, Abdul Hakeem M. ; Alraddadi, Mohammed O. ; Mahnashi, Mater H.: Skin Cancer Detection: A Review Using Deep Learning Techniques. In: *International Journal of Environmental Research and Public Health* 18 (2021), Nr. 10, S. 1–22. – URL https://doi.org/10.3390/ijerph18105479. – ISSN 1660-4601

[19] Eggert, Daniel: *Ein System für die Generierung von synthetischen fotorealistischen Bildern mit Prozessen des Deep Learning*. Hamburg, Germany, Hochschule für Angewandte Wissenschaften Hamburg, Master Thesis, May 2023. – URL https://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/master/eggert.pdf. – (Accessed 05/25/24)

[20] Foahom Gouabou, Arthur C. ; Collenne, Jules ; Monnier, Jilliana ; Iguernaissi, Rabah ; Damoiseaux, Jean-Luc ; Moudafi, Abdellatif ; Merad, Djamal: Computer Aided Diagnosis of Melanoma Using Deep Neural Networks and Game Theory: Application on Dermoscopic Images of Skin Lesions. In: *International Journal of Molecular Sciences* 23 (2022), Nr. 22, S. 1–14. – URL https://doi.org/10.3390/ijms232213838. – ISSN 1422-0067

[21] Garcea, Fabio ; Serra, Alessio ; Lamberti, Fabrizio ; Morra, Lia: Data augmentation for medical imaging: A systematic literature review. In: *Computers in Biology and Medicine* 152 (2023), S. 1–20. – URL https://doi.org/10.1016/j.compbiomed.2022.106391. – ISSN 0010-4825

[22] Gessert, Nils ; Nielsen, Maximilian ; Shaikh, Mohsin ; Werner, René ; Schlaefer, Alexander: Skin lesion classification using ensembles of multi-resolution E□cientNets with meta data. In: *MethodsX* 7 (2020), S. 1–8. – URL https://doi.org/10.1016/j.mex.2020.100864. – ISSN 2215-0161

[23] Gutman, David A. ; Codella, Noel C. F. ; Celebi, M. E. ; Helba, Brian ; Marchetti, Michael A. ; Mishra, Nabin K. ; Halpern, Allan: Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). In: *CoRR* abs/1605.01397 (2016), S. 1–5. – URL https://doi.org/10.48550/arXiv.1605.01397

[24] He, Haibo ; Garcia, Edwardo A.: Learning from Imbalanced Data. In: *IEEE Transactions on Knowledge and Data Engineering* 21 (2009), Nr. 9, S. 1263–1284. – URL https://doi.org/10.1109/TKDE.2008.239. – ISSN 1041-4347

[25] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2015, S. 1026–1034. – URL https://doi.org/10.1109/ICCV.2015.123. – ISSN 2380-7504

[26] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, S. 770–778. – URL https://doi.org/10.1109/CVPR.2016.90. – ISSN 1063-6919

[27] He, Xiaoyu ; Wang, Yong ; Zhao, Shuang ; Yao, Chunli: Deep metric attention learning for skin lesion classification in dermoscopy images. In: *Complex & Intelligent Systems* 8 (2022), Nr. 2, S. 1487–1504. – URL https://doi.org/10.1007/s40747-021-00587-4. – ISSN 2199-4536

[28] Hicks, Steven A. ; Strümke, Inga ; Thambawita, Vajira ; Hammou, Malek ; Riegler, Michael A. ; Halvorsen, Pål ; Parasa, Sravanthi: On evaluation metrics for medical applications of artificial intelligence. In: *Scientific reports* 12 (2022), Nr. 5979, S. 1–9. – URL https://doi.org/10.1038/s41598-022-09954-8

[29] Jain, Sweta ; Tripathy, Hrudaya K. ; Mallik, Saurav ; Qin, Hong ; Shaalan, Yara ; Shaalan, Khaled: Autism Detection of MRI Brain Images Using Hybrid Deep CNN With DM-Resnet Classifier. In: *IEEE Access* 11 (2023), S. 117741–117751. – URL https://doi.org/10.1109/ACCESS.2023.3325701. – ISSN 2169-3536

[30] Joseph, Seena ; Olugbara, Oludayo O.: Preprocessing E□ects on Performance of Skin Lesion Saliency Segmentation. In: *Diagnostics* 12 (2022), Nr. 2, S. 1–25. – URL https://doi.org/10.3390/diagnostics12020344. – ISSN 2075-4418

[31] Katalinic, A. ; Kunze, U. ; Schäfer, T.: Epidemiology of cutaneous melanoma and non-melanoma skin cancer in Schleswig-Holstein, Germany: incidence, clinical subtypes, tumour stages and localization (epidemiology of skin cancer). In: *British*

Journal of Dermatology 149 (2003), 12, Nr. 6, S. 1200–1206. – URL https://doi.org/10.1111/j.1365-2133.2003.05554.x. – ISSN 0007-0963

[32] KUMAR, Teerath ; MILEO, Alessandra ; BRENNAN, Rob ; BENDECHACHE, Malika: Image Data Augmentation Approaches: A Comprehensive Survey and Future directions. (2023), S. 1–32. – URL https://doi.org/10.48550/arXiv.2301.02830

[33] LEEVY, Jo□rey L. ; KHOSHGOFTAAR, Taghi M. ; BAUDER, Richard A. ; SELIYA, Naeem: A survey on addressing high-class imbalance in big data. In: *Journal of Big Data* 5 (2018), Nr. 42, S. 1–30. – URL https://doi.org/10.1186/s40537-018-0151-6

[34] LIN, Tsung-Yi ; GOYAL, Priya ; GIRSHICK, Ross ; HE, Kaiming ; DOLLÁR, Piotr: Focal Loss for Dense Object Detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020), Nr. 2, S. 318–327. – URL https://doi.org/10.1109/TPAMI.2018.2858826

[35] LIU, Lina ; MOU, Lichao ; ZHU, Xiao X. ; MANDAL, Mrinal: Automatic skin lesion classification based on mid-level feature learning. In: *Computerized Medical Imaging and Graphics* 84 (2020), S. 1–11. – URL https://doi.org/10.1016/j.compmedimag.2020.101765. – ISSN 0895-6111

[36] MAHARANA, Kiran ; MONDAL, Surajit ; NEMADE, Bhushankumar: A review: Data pre-processing and data augmentation techniques. In: *Global Transitions Proceedings* 3 (2022), Nr. 1, S. 91–99. – URL https://doi.org/10.1016/j.gltp.2022.04.020. – International Conference on Intelligent Engineering Approach (ICIEA-2022). – ISSN 2666-285X

[37] MAHBOD, Amirreza ; SCHAEFER, Gerald ; ELLINGER, Isabella ; ECKER, Rupert ; PITIOT, Alain ; WANG, Chunliang: Fusing fine-tuned deep features for skin lesion classification. In: *Computerized Medical Imaging and Graphics* 71 (2019), S. 19–29. – URL https://doi.org/10.1016/j.compmedimag.2018.10.007. – ISSN 0895-6111

[38] MAHBOD, Amirreza ; SCHAEFER, Gerald ; WANG, Chunliang ; DORFFNER, Georg ; ECKER, Rupert ; ELLINGER, Isabella: Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. In: *Computer Methods and Programs in Biomedicine* 193 (2020), S. 1–9. – URL https://doi.org/10.1016/j.cmpb.2020.105475. – ISSN 0169-2607

[39] Mahbod, Amirreza ; Schaefer, Gerald ; Wang, Chunliang ; Ecker, Rupert ; Dorffner, Georg ; Ellinger, Isabella: Investigating and Exploiting Image Resolution for Transfer Learning-based Skin Lesion Classification. In: *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, S. 4047–4053. – URL https://doi.org/10.1109/ICPR48806.2021.9412307

[40] Mehrotra, Rajat ; Ansari, M.A. ; Agrawal, Rajeev ; Anand, R.S.: A Transfer Learning approach for AI-based classification of brain tumors. In: *Machine Learning with Applications* 2 (2020), S. 1–12. – URL https://doi.org/10.1016/j.mlwa.2020.100003. – ISSN 2666-8270

[41] Mendonça, Teresa ; Ferreira, Pedro M. ; Marques, Jorge S. ; Marcal, André R. S. ; Rozeira, Jorge: PH2 - A dermoscopic image database for research and benchmarking. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2013, S. 5437–5440. – URL https://doi.org/10.1109/EMBC.2013.6610779. – ISSN 1558-4615

[42] Menzies, Scott W. ; Ingvar, Christian ; Crotty, Kerry A. ; McCarthy, William H.: Frequency and Morphologic Characteristics of Invasive Melanomas Lacking Specific Surface Microscopic Features. In: *Archives of Dermatology* 132 (1996), 10, Nr. 10, S. 1178–1182. – URL https://doi.org/10.1001/archderm.1996.03890340038007. – ISSN 0003-987X

[43] Mirikharaji, Zahra ; Abhishek, Kumar ; Bissoto, Alceu ; Barata, Catarina ; Avila, Sandra ; Valle, Eduardo ; Celebi, M. E. ; Hamarneh, Ghassan: A survey on deep learning for skin lesion segmentation. In: *Medical Image Analysis* 88 (2023), S. 1–40. – URL https://doi.org/10.1016/j.media.2023.102863. – ISSN 1361-8415

[44] Müller, Dominik ; Soto-Rey, Iñ. ; Kramer, Frank: Towards a guideline for evaluation metrics in medical image segmentation. In: *BMC research notes* 15 (2022), Nr. 210, S. 1–8. – URL https://doi.org/10.1186/s13104-022-06096-y

[45] Muller, Matthew P. ; Tomlinson, George ; Marrie, Thomas J. ; Tang, Patrick ; McGeer, Allison ; Low, Donald E. ; Detsky, Allan S. ; Gold, Wayne L.: Can Routine Laboratory Tests Discriminate between Severe Acute Respiratory Syndrome and Other Causes of Community-Acquired Pneumonia? In: *Clinical Infectious*

Diseases 40 (2005), 04, Nr. 8, S. 1079–1086. – URL https://doi.org/10.1086/428577. – ISSN 1058-4838

[46] Nahm, Francis S.: Receiver operating characteristic curve: overview and practical use for clinicians. In: *Korean J Anesthesiol* 75 (2022), Nr. 1, S. 25–36. – URL https://doi.org/10.4097/kja.21209

[47] Nida, Nudrat ; Irtaza, Aun ; Javed, Ali ; Yousaf, Muhammad H. ; Mahmood, Muhammad T.: Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering. In: *International Journal of Medical Informatics* 124 (2019), S. 37–48. – URL https://doi.org/10.1016/j.ijmedinf.2019.01.005. – ISSN 1386-5056

[48] Nie, Yali ; Sommella, Paolo ; Carratù, Marco ; O'Nils, Mattias ; Lundgren, Jan: A Deep CNN Transformer Hybrid Model for Skin Lesion Classification of Dermoscopic Images Using Focal Loss. In: *Diagnostics* 13 (2023), Nr. 1, S. 1–18. – URL https://doi.org/10.3390/diagnostics13010072. – ISSN 2075-4418

[49] Nischal, Urmila ; Nischal, Kc ; Khopkar, Uday: Techniques of skin biopsy and practical considerations. In: *Journal of cutaneous and aesthetic surgery* 1 (2008), Nr. 2, S. 107–111. – URL https://doi.org/10.4103/0974-2077.44174

[50] Patel, Hima ; Yacoub, Nour ; Mishra, Rosalin ; White, Aaron ; Yuan, Long ; Alanazi, Samar ; Garrett, Joan T.: Current Advances in the Treatment of BRAF-Mutant Melanoma. In: *Cancers* 12 (2020), Nr. 2, S. 1–44. – URL https://doi.org/10.3390/cancers12020482. – ISSN 2072-6694

[51] Pewton, Samuel W. ; Yap, Moi H.: Dark Corner on Skin Lesion Image Dataset: Does it matter? In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2022, S. 4830–4838. – URL https://doi.org/10.1109/CVPRW56347.2022.00530. – ISSN 2160-7516

[52] Phung, Van H. ; Rhee, Eun J.: A Deep Learning Approach for Classification of Cloud Image Patches on Small Datasets. In: *Journal of Information and Communication Convergence Engineering* 16 (2018), 9, Nr. 3, S. 173–178. – URL https://doi.org/10.6109/jicce.2018.16.3.173

[53] Rezk, Eman ; Eltorki, Mohamed ; El-Dakhakhni, Wael: Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach. In: *JMIR Dermatol* 5

(2022), Aug, Nr. 3, S. 1–14. – URL https://doi.org/10.2196/39143. – ISSN 2562-0959

[54] Saba, Tanzila ; Khan, Muhammad A. ; Rehman, Amjad ; Marie-Sainte, Souad L.: Region Extraction and Classification of Skin Cancer: A Heterogeneous framework of Deep CNN Features Fusion and Reduction. In: *Journal of medical systems* 43 (2019), Nr. 289, S. 1–19. – URL https://doi.org/10.1007/s10916-019-1413-3

[55] Sabottke, Carl F. ; Spieler, Bradley M.: The E□ect of Image Resolution on Deep Learning in Radiography. In: *Radiology. Artificial intelligence* 2 (2020), Nr. 1, S. 1–7. – URL https://doi.org/10.1148/ryai.2019190015

[56] Saeed, Jwan ; Zeebaree, Subhi: Skin Lesion Classification Based on Deep Convolutional Neural Networks Architectures. In: *Journal of Applied Science and Technology Trends* 2 (2021), March, Nr. 01, S. 41–51. – URL https://doi.org/10.38094/jastt20189

[57] Saito, Takaya ; Rehmsmeier, Marc: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. In: *PLOS ONE* 10 (2015), 03, Nr. 3, S. 1–21. – URL https://doi.org/10.1371/journal.pone.0118432

[58] Sarwinda, Devvi ; Paradisa, Radifa H. ; Bustamam, Alhadi ; Anggia, Pinkie: Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. In: *Procedia Computer Science* 179 (2021), S. 423–431. – URL https://doi.org/10.1016/j.procs.2021.01.025. – 5th International Conference on Computer Science and Computational Intelligence 2020. – ISSN 1877-0509

[59] Scolyer, Richard A. ; Rawson, Robert V. ; Gershenwald, Je□rey E. ; Ferguson, Peter M. ; Prieto, Victor G.: Melanoma pathology reporting and staging. In: *Modern Pathology* 33 (2020), Nr. 1, S. 15–24. – URL https://doi.org/10.1038/s41379-019-0402-x. – ISSN 1530-0285

[60] Shetty, Bhuvaneshwari ; Fernandes, Roshan ; Rodrigues, Anisha P. ; Chengoden, Rajeswari ; Bhattacharya, Sweta ; Lakshmanna, Kuruva: Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. In: *Scientific reports* 12 (2022), Nr. 18134, S. 1–11. – URL https://doi.org/10.1038/s41598-022-22644-9

[61] Shorten, Connor ; Khoshgoftaar, Taghi M.: A survey on Image Data Augmentation for Deep Learning. In: *Journal of Big Data* 6 (2019), Nr. 60, S. 1–48. – URL https://doi.org/10.1186/s40537-019-0197-0

[62] Showkat, Sadia ; Qureshi, Shaima: E□cacy of Transfer Learning-based ResNet models in Chest X-ray image classification for detecting COVID-19 Pneumonia. In: *Chemometrics and Intelligent Laboratory Systems* 224 (2022), S. 1–10. – URL https://doi.org/10.1016/j.chemolab.2022.104534. – ISSN 0169-7439

[63] Sies, Katharina ; Winkler, Julia K. ; Fink, Christine ; Bardehle, Felicitas ; Toberer, Ferdinand ; Kommoss, Felix K. F. ; Buhl, Timo ; Enk, Alexander ; Rosenberger, Albert ; Haenssle, Holger A.: Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification. In: *JDDG: Journal der Deutschen Dermatologischen Gesellschaft* 19 (2021), Nr. 6, S. 842–850. – URL https://doi.org/10.1111/ddg.14384

[64] Sreelatha, Tammineni ; Subramanyam, M. V. ; Prasad, M. N. G.: Early Detection of Skin Cancer Using Melanoma Segmentation technique. In: *Journal of medical systems* 43 (2019), Nr. 190, S. 1–7. – URL https://doi.org/10.1007/s10916-019-1334-1

[65] Srivastava, Nitish ; Hinton, Geo□rey ; Krizhevsky, Alex ; Sutskever, Ilya ; Salakhutdinov, Ruslan: Dropout: a simple way to prevent neural networks from overfitting. In: *The Journal of Machine Learning Research* 15 (2014), January, Nr. 1, S. 1929–1958. – URL https://dl.acm.org/doi/10.5555/2627435.2670313. – ISSN 1532-4435

[66] Tang, Shuzhen ; Jing, Chen ; Jiang, Yitao ; Yang, Keen ; Huang, Zhibin ; Wu, Huaiyu ; Cui, Chen ; Shi, Siyuan ; Ye, Xiuqin ; Tian, Hongtian ; Song, Di ; Xu, Jinfeng ; Dong, Fajin: The e□ect of image resolution on convolutional neural networks in breast ultrasound. In: *Heliyon* 9 (2023), Nr. 8, S. 1–7. – URL https://doi.org/10.1016/j.heliyon.2023.e19253. – ISSN 2405-8440

[67] Thambawita, Vajira ; Strümke, Inga ; Hicks, Steven A. ; Halvorsen, Pål ; Parasa, Sravanthi ; Riegler, Michael A.: Impact of Image Resolution on Deep Learning Performance in Endoscopy Image Classification: An Experimental Study Using a Large Dataset of Endoscopic Images. In: *Diagnostics* 11 (2021), Nr. 12, S. 1–9. – URL https://doi.org/10.3390/diagnostics11122183. – ISSN 2075-4418

[68] Tschandl, Philipp ; Rosendahl, Cli ; Kittler, Harald: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. In: *Scientific Data* 5 (2018), Nr. 180161, S. 1–9. – URL https://doi.org/10.1038/sdata.2018.161. – ISSN 2052-4463

[69] Unlu, Ezgi ; Akay, Bengu N. ; Erdem, Cengizhan: Comparison of dermatoscopic diagnostic algorithms based on calculation: The ABCD rule of dermatoscopy, the seven-point checklist, the three-point checklist and the CASH algorithm in dermatoscopic evaluation of melanocytic lesions. In: *The Journal of Dermatology* 41 (2014), Nr. 7, S. 598–603. – URL https://doi.org/10.1111/1346-8138.12491

[70] Venugopal, Vipin ; Nath, Malaya K. ; Joseph, Justin ; Das, M. V.: A deep learning-based illumination transform for devignetting photographs of dermatological lesions. In: *Image and Vision Computing* 142 (2024), S. 1–20. – URL https://doi.org/10.1016/j.imavis.2024.104909. – ISSN 0262-8856

[71] Vincent, Luc: Morphological transformations of binary images with arbitrary structuring elements. In: *Signal Processing* 22 (1991), Nr. 1, S. 3–23. – URL https://doi.org/10.1016/0165-1684(91)90025-E. – ISSN 0165-1684

[72] Vocaturo, Eugenio ; Zumpano, Ester ; Veltri, Pierangelo: Image pre-processing in computer vision systems for melanoma detection. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, S. 2117–2124. – URL https://doi.org/10.1109/BIBM.2018.8621507

[73] Wang, Xiaosong ; Peng, Yifan ; Lu, Le ; Lu, Zhiyong ; Bagheri, Mohammadhadi ; Summers, Ronald M.: ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, S. 3462–3471. – URL https://doi.org/10.1109/CVPR.2017.369

[74] Wang, Ying ; Su, Jie ; Xu, Qiuyu ; Zhong, Yixin: A Collaborative Learning Model for Skin Lesion Segmentation and Classification. In: *Diagnostics* 13 (2023), Nr. 5, S. 1–16. – URL https://doi.org/10.3390/diagnostics13050912. – ISSN 2075-4418

[75] World Health Organisation: *Cancer*. Feb 2022. – URL https://www.who.int/news-room/fact-sheets/detail/cancer. – (Accessed 04/10/24)

[76] Xu, Chuhan ; Coen-Pirani, Pablo ; Jiang, Xia: Empirical Study of Overfitting in Deep Learning for Predicting Breast Cancer Metastasis. In: *Cancers* 15 (2023), Nr. 7, S. 1–18. – URL https://doi.org/10.3390/cancers15071969. – ISSN 2072-6694

[77] Yuan, Yading ; Chao, Ming ; Lo, Yeh-Chi: Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance. In: *IEEE Transactions on Medical Imaging* 36 (2017), Nr. 9, S. 1876–1886. – URL https://doi.org/10.1109/TMI.2017.2695227

[78] Zafar, Kashan ; Gilani, Syed O. ; Waris, Asim ; Ahmed, Ali ; Jamil, Mohsin ; Khan, Muhammad N. ; Sohail Kashif, Amer: Skin Lesion Segmentation from Dermoscopic Images Using Convolutional Neural Network. In: *Sensors* 20 (2020), Nr. 6, S. 1–14. – URL https://doi.org/10.3390/s20061601. – ISSN 1424-8220

[79] Zhuang, Fuzhen ; Qi, Zhiyuan ; Duan, Keyu ; Xi, Dongbo ; Zhu, Yongchun ; Zhu, Hengshu ; Xiong, Hui ; He, Qing: A Comprehensive Survey on Transfer Learning. In: *Proceedings of the IEEE* 109 (2021), Nr. 1, S. 43–76. – URL https://doi.org/10.1109/JPROC.2020.3004555

Erklärung zur selbständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit „— bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] — ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen."

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Erklärung zur selbständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: _____

Vorname: _____

dass ich die vorliegende Bachelorarbeit – bzw. bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit – mit dem Thema:

Effects of Image Resolution on Skin Lesions Classification with Residual Neural Networks

ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

_____	_____	_____
Ort	Datum	Unterschrift im Original