Ernst-Abbe-Hochschule Jena
University of Applied Sciences

Department of Electrical Engineering and Information Technology

Master Thesis

to obtain the grade of

Master of Engineering (M.Eng.):

Domain Shifts and Interpretability in AI-based Skin Cancer Diagnosis

Domänenverschiebung und Erklärbarkeit in KI-basierter Hautkrebserkennung

Submitted by:	Markus Till Ertmer
Date of birth:	July, 7, 1997
Student number :	643417
Study program:	Elektrotechnik/Informationstechnik
Academic supervisor:	Prof. Dr. Sebastian Knorr
Supervisor:	Dr. Sireesha Chamarthi
Starting date:	October, 5, 2023
Submission date:	October, 19, 2023

Jena, October 2023

List of Contents

List of Figures

List of Tables

List of Abbreviations

CNN	Convolutional Neural Network
ISIC	International Skin Imaging Collaboration
ResNet	Residual Network
Grad-CAM	Gradient-weighted Class Activation Mapping
HiResCAM	High-Resolution Class Activation Mapping
MRI	Magnetic Resonance Imaging
DNN	Deep neural network
DANN	Domain-Adversarial Training of Neural Networks
AUROC	Area Under the Receiver Operating Characteristic curve
IoU	Intersection over Union
wIoU	Weighted Intersection over Union
BCDU-Net	Bi-Directional ConvLSTM U-Net with Densely Connected Convolutions
ReLU	Rectified Linear Units
CAM	Class Activation Map
t-SNE	t-Distributed Stochastic Neighbor Embedding
SVHN	Street View House Numbers
ConvLSTM	Convolutional Long Short-Term Memory
SGD	Stochastic Gradient Descent
HAM	Human Against Machine
CSV	Comma-Separated Values
RGB	Red, Green and Blue
BCN	Barcelona (Hospital Clínic)
MSK	Memorial Sloan-Kettering (Cancer Center)
UDA	Unsupervised Domain Adaptation
COAL	Feature and Label distribution Co-Alignment Model

Abstract

Explainability is crucial in order to build trust in AI applications and increase their acceptance, particularly in safety-critical environments. One such application is skin cancer classification, where dermatologists may use AI models as digital assistants for diagnostic purposes. In this work, the publicly available domain-separated ISIC-Archive dataset, comprised of melanoma and nevus image data, is examined. Initially, an inter-domain dataset is used to train a binary ResNet18 classifier and the Grad-CAM output is interpreted. Subsequently, the performance and Grad-CAM output of a binary classifier is investigated to understand how neural network activations change during domain shifts within the ISIC-Dataset. Thereafter, the established domain shift mitigation approaches - i.e. augmentation and DANN - are investigated regarding their influence on performance and neural network activations. Their effects on Grad-CAM-Elementwise output are furthermore quantitatively compared. The findings presented in this work, provide insights into the underlying reasons for unequal performance degradation during domain shift, shortcomings of Grad-CAM and influential factors of unsupervised domain adaptation. The results indicate a limited potential of Grad-CAM based explainability methods for building trust among dermatologists in domain-adapted ResNet models. Nevertheless, Grad-CAM based explainability methods have demonstrated their ability to identify failure modes of neural networks.

Abstract

Erklärbarkeit ist entscheidend, um Vertrauen und Akzeptanz in KI-Anwendungen aufzubauen, insbesondere in sicherheitskritischen Umgebungen. Eine solche Anwendung ist die Klassifikation von Hautkrebs, bei der Dermatologen KI-Modelle als digitale Assistenten diagnostische Zwecke nutzen und deren Ergebnisse nachvollziehen möchten. In dieser Arbeit wird der öffentlich verfügbare, nach Domänen getrennte ISIC-Archive-Datensatz, der aus Bildern von Melanomen und Nävi besteht, untersucht. Zunächst wird ein mehrere Domänen umfassender Datensatz verwendet, um einen binären ResNet18-Klassifikator zu trainieren. Die Grad-CAM-Ausgabe wird anschließend interpretiert. Danach werden die Leistungsmerkmale und die Grad-CAM-Ausgaben eines binären Klassifikators untersucht, um zu verstehen, wie sich Aktivierungen des neuronalen Netzwerks während der Domänenwechsel im ISIC-Datensatz ändern. Im weiteren Verlauf werden zwei etablierte Ansätze zur Minderung der Domänenverschiebung - Datenmanipulation und DANN - nach ihrem Einfluss auf Leistungsmerkmale und Aktivierungen des neuronalen Netzwerks untersucht. Ihre Auswirkungen auf die Grad-CAM-Elementwise Ausgabe werden außerdem quantitativ verglichen. Die in dieser Arbeit präsentierten Ergebnisse liefern Einblicke in die Gründe für die ungleiche Leistungsverschlechterung während der Domänenverschiebung, in verschiedene Unzulänglichkeiten von Grad-CAM und in die Einflussfaktoren der unüberwachten Domänenanpassung. Die Ergebnisse deuten auf ein begrenztes Potenzial von Grad-CAM-basierten Erklärbarkeitsmethoden für die Schaffung von Vertrauen in domänenadaptierte ResNet-Modelle hin. Nichtsdestotrotz haben Grad-CAM-basierte Erklärbarkeitsmethoden ihre Eignung zur Identifizierung von Fehlermodi neuraler Netze unter Beweis gestellt.

1. Introduction

Skin cancer poses a global health problem, with malignant melanoma (black skin cancer) standing out as the most lethal form within this category. The early diagnosis of malignant melanoma warrants the chance for complete recovery of the patient [1]. Distinguishing between malignant melanoma and a benign melanocytic nevus (mole) poses a considerable challenge. Therefore, the expertise of a proficient dermatologist is required for early diagnosis. However, the accuracy of the diagnosis is influenced by the experience of the involved dermatologist [2], [3]. Given the rising incidence of melanoma across many countries [4], reliable computer vision solutions promise the potential to assist in saving many lives.

In addition to various patient specific factors, dermoscopic images represent the primary source of information for a dermatologist to eventually validate the malignance of skin lesions such as moles. These images allow the dermatologist to assess the microstructures of a patient's epidermis [5]. Therefore, it appears logical and beneficial to train ai based models using comparable data that dermatologists utilize to formulate their diagnoses and determine whether to excise or retain a skin lesion. This approach theoretically enables software implementations to aid dermatologists in enhancing their diagnostic precision, leading to an overall improvement in accuracy.

The broad adoption of AI-based skin cancer classification is held back by the limited generalization capability of skin cancer classification models [6]. Within the available dermoscopic image data, the generalization capabilities of Convolutional Neural Networks (CNNs) are tested on new domains. Consequently, a CNN capable of maintaining performance on the test dataset of another domain is considered to possess robust generalization capabilities. Conclusively, domain adaptation in skin cancer classification is a key aspect in the efforts to convey AI-based skin cancer classification into the clinic. Since CNNs inherently operate as black boxes, the underlying mechanisms of influential factors, leading to lower performance upon domain shift, remain unknown. Another crucial aspect for application in clinical diagnosis is the explainability of classifiers and their reliability. Establishing trust among dermatologists is essential for a wide use of AI-based software assistants.

1.1 Objectives

To gain further insights into influential factors of domain shift and the behavior of explainability methods in domain shift scenarios, the following objectives are formulated.

Neural network activations in the International Skin Imaging Collaboration (ISIC) -Dataset are analyzed during domain shifts to understand the causes of domain shift. Additionally, the changes of a neural network's activation are examined as the domain is changed within the ISIC-Dataset, to observe generalization capabilities.

After identifying the underlying reasons for domain shift, established mitigation approaches are investigated. The first approach is the augmentation of dermoscopic images. Its impact is investigated on both the classification performance and neural network activations.

Subsequently, the most suitable saliency method for quantifying the impact of augmentation and domain adaptation on network activations is determined.

This is followed by domain adaptation and the investigation of its impact on both the classification performance and neural network activations

The influence of mitigation approaches on neural network activation is then quantitatively compared, to gain insights in the explainability of domain adapted models.

2. Related Work

In this chapter existing literature and relevant research is discussed to illustrate how the key points and employed methods of this work are integrated in research.

2.1 Research on Dermoscopic Skin Cancer Classification

A prominent initial publication addressing skin cancer classification with AI-based techniques was published in 2017 [7]. This study introduced the utilization of a pretrained GoogleNet Inception v3 CNN architecture for distinguishing nine different skin disease classes. Remarkably, the model's performance was on par with that of a dermatologist. To accelerate the ambition of enhanced achievements in classifying dermoscopic skin lesions, the ISIC-Challenge [8] was established. Since a wide variety of Networks was deployed in the ISIC-Challenges, efforts emerged to determine more suitable CNN architectures for this task. This led up to the development of ensemble learning [9], wherein multiple models are applied and their outcomes consolidated to obtain the final classification result.

Other bounding aspects as dataset imbalance, the limited amount of labeled data in skin disease datasets and the inadequate cross-domain generalization capabilities of trained models have also been subject of exploration within publications [3].

Recent publications target the aforementioned aspects and try to contribute with an enhancement in one of these aspects [10]–[14].

Residual Networks (ResNets) have found application in the realm of dermoscopic classification, as demonstrated in [15] and particularly have been investigated and compared to other CNNs for their transfer learning capabilities in skin cancer classification with convincing results [16].

2.2 Saliency as Explainability and the Role of Grad-CAM

In implementing CNNs in safety critical tasks like skin lesion classifications, the acceptance of the models among the medical community depends on explaining the model's decisions. In the case of image classification tasks, it is advisable to consider a visualization strategy that mirrors human assessment methods, similar to those used in generating saliency maps [17]. Early publications retrieved the pixel importance solely by backpropagation through the CNN [18]. The success of Grad-CAM [19] as a saliency method led to the emergence of several other approaches based on Grad-CAM such as HiResCAM [20], Grad-CAM++ [21] and FullGrad [22], to name just a few instances.

Since its publication, Grad-CAM has found extensive usage in the scientific field. For instance, it was employed to highlight areas of brain MRI-scans [23]. Beyond its medical applications, Grad-CAM was also applied in explaining domain adaptation and exploring generalization capabilities of DNNs while directly factoring saliency output into the model’s training [24].

Grad-CAM has additionally found application in quantification tasks, assessing similarity between human saliency and neural network attention [25]. The quantification metrics employed are those proposed by Boggust et al. [26], some of which will be also used in this work.

2.3 Domain Adaptation in Dermoscopic Datasets (Augmentation vs DANN)

The most relevant study is a publication [10], which identified and quantified domain shifts within the ISIC-Dataset. This dataset, comprising separated domains and quantified domain shifts, builds a foundation for this work. The research further suggests that employing techniques like data augmentation or domain adaptation may offer potential solutions for the issue of domain shifts. The authors focused on unsupervised domain adaptation to demonstrate that the effects of the identified domain shifts can be impaired. The architecture they employed featured a ResNet50 model, utilized as a feature extractor within the framework of DANN. By evaluating DANN’s AUROC performance across the identified domains, the study highlighted DANN’s capability to enhance performance across the majority of domains [10].

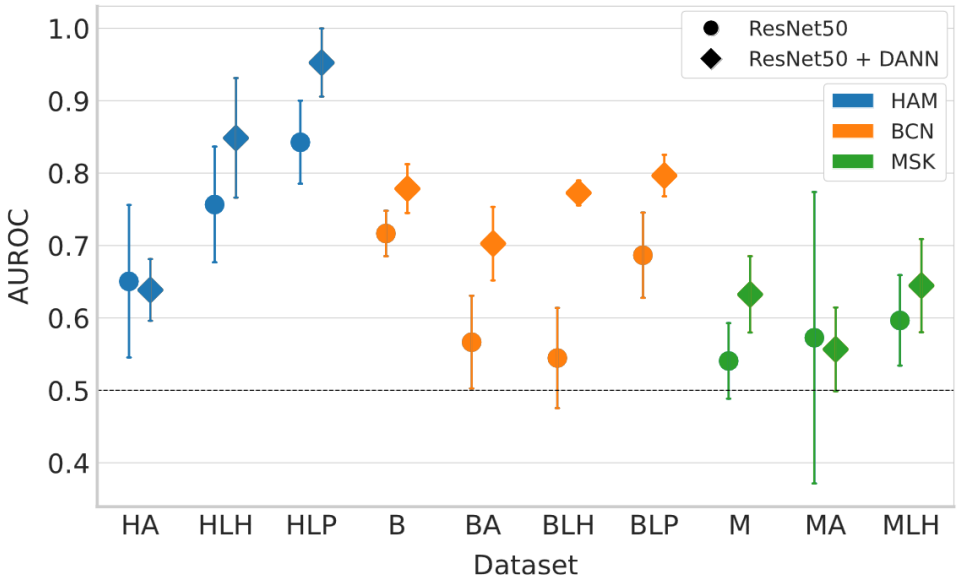

Figure 1: Effect of DANN in combination with ResNet50 on classification performance across domains inside the ISIC-Dataset [10]

While the authors could identify the domains in the data structure based on meta-data, they did not provide a definitive explanation of these shifts on image level. This work aims to contribute to this particular research field by addressing this gap.

2.4 Segmentation in AI-based Skin Cancer Classification

Segmentation represents a valuable tool for quantification purposes. The 2018 ISIC challenge [8] evaluated dermoscopic segmentation algorithms based on labeled training images and masks. For this challenge a dataset with 2594 images was provided with the according segmentation masks which were created by medical professionals. The leading submission achieved 0.82 Intersection over Union (IoU) on the test set. Subsequently, various attempts tried to further optimize performance on this task, like Bi-Directional ConvLSTM U-Net with Densely Connected Convolutions (BCDU-Net) [27]. With a Jaccard-Score (IoU) of 0.937 on a comparable test-dataset BCDU-Net surpasses the performances of the earlier proposed U-Net and the ISIC-Challenge winner.

It is important to highlight that the ground truth masks occasionally contain linear borders. Given the scarceness of linearly delimited skin lesions, one could argue that the ground truth sometimes wasn't prepared with the necessary conscientiousness.

This leaves room for interpretation. In this context, BCDU-Net segmentation masks might be regarded to represent the ground truth. Based on this assumption, the skin lesion segmentation masks generated by BCDU-Net could be employed for quantifying Grad-CAM output. In addition, a quantitative comparison of Grad-CAM and statistical results becomes possible.

3. Technical Background

In this chapter relevant technical fundamentals are presented, which are required for comprehension of the later presented methods and appraising the results.

3.1 The ResNet Architecture

The ResNet architecture [28] is primarily applied in CNNs, dedicated to image recognition. It was proposed in 2015 and since then found great resonance in the scientific community [29]. The different ResNet models are available in several levels of complexity. Today, networks between 18 and 152 layers are available inside the PyTorch framework. In skin cancer classification there have been various uses of this architecture to process image data [10], [30], [31].

The main problem this residual approach addresses is the vanishing gradient problem, which was first described by Hochreiter [32]. During backpropagation in deep neural networks this problem leads to a gradient which approaches zero for the early layers of the model. Therefore, the model is learning slowly and eventually does not exhibit adequate test results.

One layer of a ResNet18 model consists of two building blocks. Each block receives an input x that is introduced to a weight layer (see Figure 2). The Rectified Linear Unit (ReLU) function is applied along with repeated weighting. Finally, the initial input is added to the output of the second weight layer [28]. This shortcut connection enables gradients to directly flow backwards to the early layers of the model during backpropagation [33].

This leads to ResNet models exhibiting superior performance compared to traditional plain CNNs with same size and same number of training iterations [28].

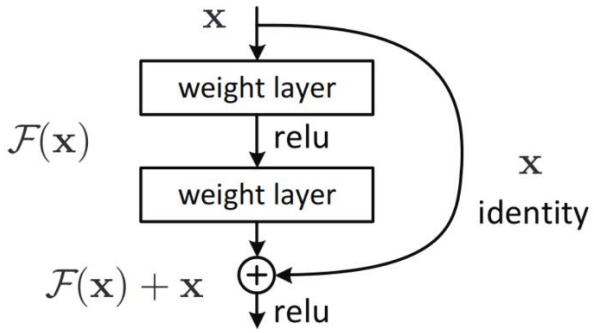

Figure 2:ResNet architecture building block [28]

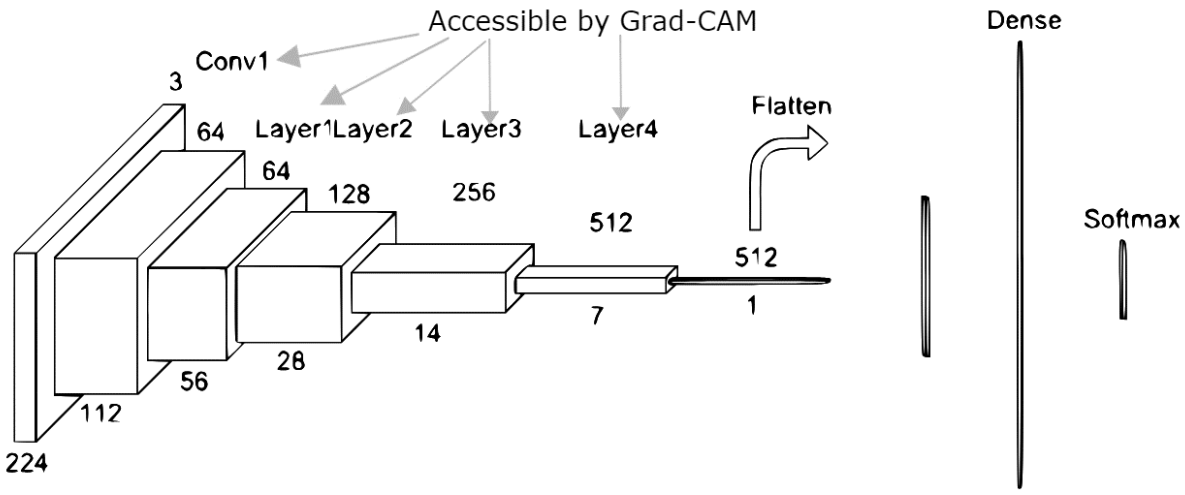

Figure 3:ResNet architecture with layers that contain building blocks [33]

In this work, a pretrained ResNet18 model is being used as a classifier and feature extractor. The ResNet18 model has 18 layers which form 5 Blocks (see Figure 3). These 5 blocks consist of a different number of layers depending on the chosen ResNet model. Dimensionality is maintained inside each block, because the padding is 1, kernel width and height is 3. The output size is determined by equation 1.

$$O_{size} = \frac{I_{size}+2*P-K}{S} + 1 \rightarrow O_{size} = \frac{I_{size}+2-3}{1} + 1 = I_{size}\ [34] \quad (1)$$

O_{size} = output size; I_{size} = input size; P= padding; K = kernel size; S = stride

Most convolutional layers of the ResNet architecture contribute to the extraction of the relation of the neighboring features. However, the characteristic shortcuts between the layers make the learning process significantly more effective with ResNet models of higher complexity. Even though this benefit is marginal when using the ResNet18 model, it offers comparably solid results with a low number of operations.

To increase effectiveness a Transfer Learning approach, that is available inside the PyTorch framework, is used for training the ResNet18 model. More specifically, a pretrained ResNet model can be loaded with weights that were obtained by training on the ImageNet-1K dataset.

3.2 Saliency Methods

3.2.1 Grad-CAM

The Gradient-weighted Class Activation Mapping aims to create pixel intensity maps that can highlight failure modes, dataset bias and class specific features of CNNs [19]. Failure modes simply describe in what way a CNN fails. In concrete terms, Grad-CAM creates a saliency map for a CNN, later also referred to as Class Activation Map (CAM) or simply activation. The Grad-CAM visual explanation technique was proposed in 2016. For this work, the most popular

GitHub [1]Grad-CAM implementation by Jacob Gildenblat [35] was used. The implementation features several different diversifications of the original Grad-CAM.Grad-CAM can only be applied to networks that contain convolutional layers, since only convolutional layers retain spatial information which is lost in fully connected layers. Grad-CAM is usually applied to the last convolutional layer of a model, since this layer contains high level features and usually highlights cohesive regions in the image. When applied to the last convolutional layer, Grad-CAM uses the gradient information that is fed into the last convolutional layer of the CNN [19].

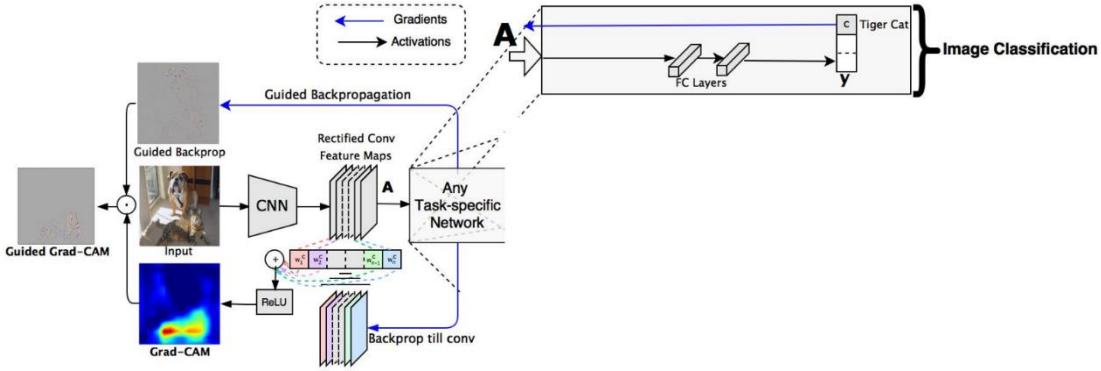

Figure 4:Grad-CAM principle [19]

As visible in the schematic in Figure 4, an image is fed forward through the network and the corresponding class is assigned. As shown in Figure 4, Grad-CAM sets the gradient of the target class "Tiger Cat" to 1 and the other classes are forced to 0. This is followed by a backpropagation through the network until the chosen target layer. At this target layer the class-discriminative localization map is computed.

First the neuron importance weights α for the target class c and each feature map k are calculated. Summing over height and width of the feature map and dividing by the number of features is also referred to as global average pooling. This results in an average gradient for every feature map in the target layer [34].

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad [19] \qquad (2)$$

Whereas i and j are height and width of the feature map and Z the number of features. A represents the matrix of the feature map activations.

[1] https://github.com/jacobgil/pytorch-grad-cam

Second, the class-discriminative localization map is computed by multiplying the average gradient with the feature map activations before summing over all feature maps [19].

$$L^c_{GradCAM} = ReLU(\sum_k \alpha^c_k A^k) \quad [19] \qquad (3.1)$$

The ReLU ensures that the values of $L^c_{GradCAM}$ are positive [19]

3.2.2 Grad-CAM-Elementwise

The Grad-CAM Git-Hub repository [2]offers various other saliency methods. One of these is Grad-CAM-Elementwise [35], which essentially represents an extension of HiResCAM, which in itself is an extension of Grad-CAM.

In the following section, HiResCAM and its difference to Grad-CAM shall be outlined. The functional principle is similar to Grad-CAM, but the feature map activations and the gradients are multiplied elementwise. The main concern about Grad-CAM, that the HiResCAM publication expounds is the use of the average gradient, that is multiplied with the feature map activations. The authors visualized the problem in Figure 5 [20].

It is evident that Grad-CAM relies more on the feature map values in its final representation.

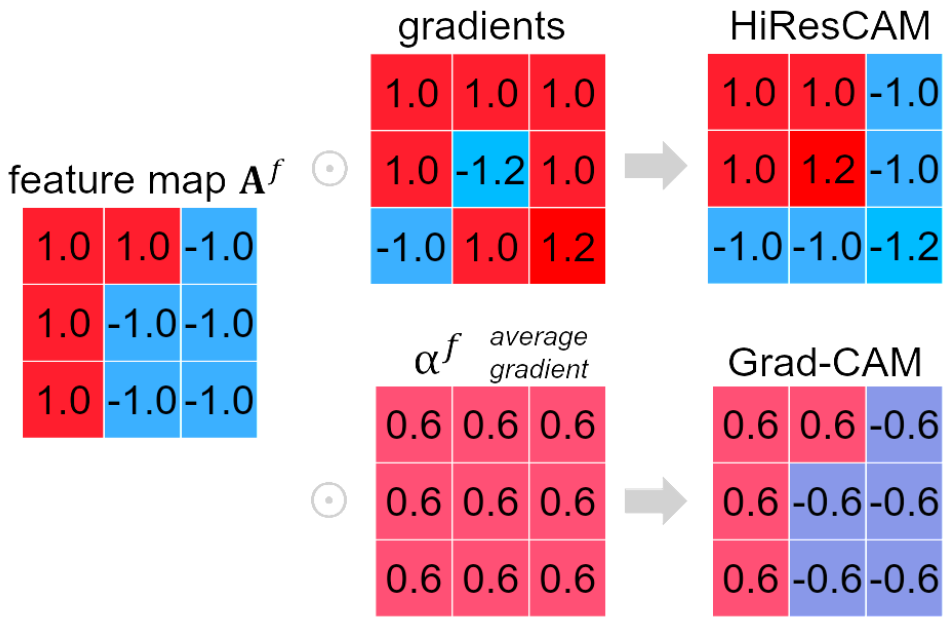

Figure 5: Difference between HiResCAM and Grad-CAM in 2D perspective [20]

[2] https://github.com/jacobgil/pytorch-grad-cam

Mathematically, HiResCAM is described as a matrix of the gradient of the score y for class c with respect to feature map activations elementwise multiplied with the feature map matrix [20].

$$L^c_{HiResCAM} = \sum_k \frac{\partial y^c}{\partial A^k} \odot A^k \quad [20] \tag{3.2}$$

Grad-CAM-Elementwise is only an additional Saliency Method implementation in the Grad-CAM GitHub repository. Since it is described as an elementwise multiplication of the activations with the gradients to which ReLU is applied before summing [35]. It is Mathematically seen as:

$$L^c_{GradCAMElementWise} = \sum_k ReLU\left(\frac{\partial y^c}{\partial A^k} \odot A^k\right) \tag{3.3}$$

As a result of ReLU, the product of an element is negative it gets set to zero. While summing over all the feature maps inside the target layer, those negative values are blocked from contributing to the result. Consequently, feature maps or gradient matrices with mostly negative gradients have low impact on the final class-discriminative localization map.

If the feature map A^f in Figure 5 and the corresponding gradients are multiplied in the way that equation 3.3 specifies it, the result would be alike the matrix in Figure 6.

GradCAMElementWise

1.0	1.0	0
1.0	1.2	0
0	0	0

Figure 6: Respective 2D Result for Grad-CAM-Elementwise (created with draw.io)

3.3 Augmentation of Datasets in Machine Learning

Data augmentation is a prevalent technique in machine learning aimed at mitigating overfitting during the training of models [36]. In the context of image data, various augmentation methods are employed, comprising geometric transformations, color space conversions, kernel-based filters, image erasure, and image blending. Additionally, synthetic data generation by generative models can prove useful when dealing with datasets primarily composed of natural images [37]. Data augmentation has already found utility within medical image datasets, particularly for achieving domain adaptation objectives. Beyond the scope of geometric augmentation, dataset specific augmentation techniques [38], [39] have been explored in various works. This specifically also features the removal and adding of artifacts. Data

augmentation positively contributes to the robustness of trained models, consequently facilitating improved generalization performance. Notably, Tesla employs dataset augmentation to increase the robustness of their Autopilot systems with a patented approach [40], [41]. While data augmentation not inherently being a domain adaptation method, it is evident to hypothesize that models showcasing notable generalization capacity also exhibit an ability to perform well on unseen domains.

3.4 Domain Adaptation with DANN

In the realm of machine learning, domain adaptation is a sub-field that aims at aligning the discrepancy across the domain data distribution to generalize a trained model into the domain of interest [42]. In a broader sense, domain adaptation is also a special variant of transfer learning, since already gathered knowledge is used to enhance performance on a closely related task.

For domain adaptation there are at least two domains that contain a discrepancy in their data, a source domain and a target domain. Usually, the source domain contains labeled data that a model was initially trained on. Domains are comprised of three components. A feature space, label space and their associated probability distribution. Feature space is a subset of a multidimensional space [42].

In case of image data, this data is equivalent to the entirety of the feature maps of one layer of a CNN e.g. ResNet. In case of ResNet18, a vector with the size 1x1x512 is retrieved after average pooling [28]. Statistical methods like t-Distributed Stochastic Neighbor Embedding (t-SNE) can reduce the dimensionality of this vector. By that, the information contained inside an image is represented as a coordinate in a two-dimensional feature space, as shown in Figure 7.

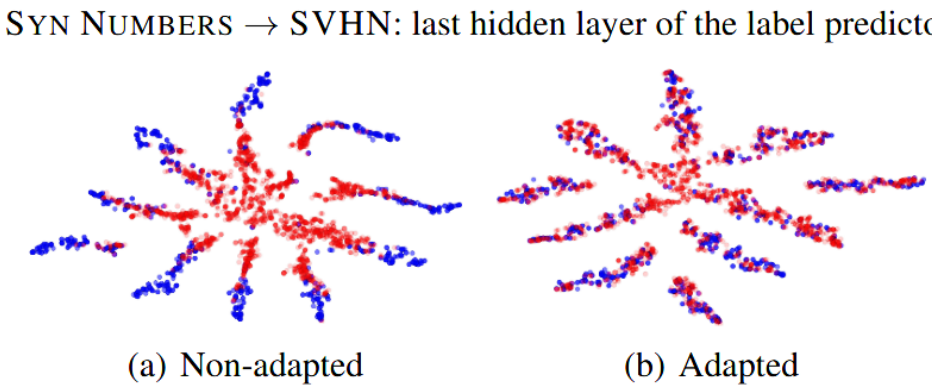

Figure 7: Domain adaptation effect of the trained feature extractor on the image data. Here represented by t-SNE visualization of the feature extractors output. Blue dots represent SYN NUMBERS, red dots represent SVHN samples. [43]

The label space is comprised of the number of classes contained inside the domain. The probability distribution describes the probability that an instance inside the data, e.g. an image, has a label y when it is located at point x somewhere in the feature space [42].

Domain shift describes the change of the data distribution between the source- and target domain [44]. For this work, only the type of unsupervised domain adaptation plays a role. Here the source domain contains labeled training data whereas the data inside the target domain is unlabeled.

In Figure 7, the effect of DANN on the data distribution can be observed with the Synthetic Digits (SYN Numbers) and the Street View House Numbers (SVHN) datasets. After domain adaptation, images of the target domain share a similar feature space to the source domain images. Thus, they should be easier to identify with knowledge that is restricted to the source domain.

DANN, which was used to realize the domain adaptation in this work, fits inside the scope of deep domain adaptation and is one of the most well established domain adaptation approaches[45].

DANN has already been used in domain adaptation for skin cancer classification and demonstrated promising results [10].

The fundamental idea behind DANN is to align the feature spaces of the source and target domains. If the features from both domains are indistinguishable, they can be considered similar or equal. As a consequence, a classifier that bases its prediction on source domain features should exhibit comparable performance for source and target domain. The functional principal is outlined in Figure 8 [43].

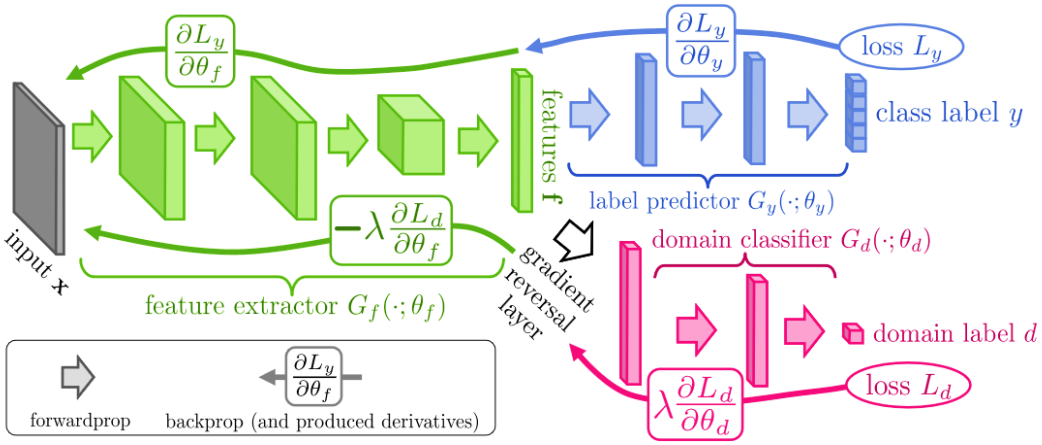

Figure 8: Proposed DANN architecture [43]

The mentioned features are derived from the abstraction of image data. Since the application is image classification, the label predictor aims to classify the image based on the features, that the feature extractor, e.g. ResNet, outputs. Additionally, the output of the feature extractor is passed to the domain classifier. The domain classifier distinguishes to which domain the features correspond to. Since the architectural aim here is to bring the feature spaces of the two domains into alignment, a gradient reversal layer is introduced between the feature extractor and the domain classifier. The function of the gradient reversal layer is to invert the gradient associated with the loss of the domain classifier during the backward propagation phase [43]. Consecutively, the feature extractor is influenced to optimize its parameters according to the reversed domain classifier loss and the label predictor loss. In concrete terms, a correct label prediction and an incorrect domain classification of image features generates minimal loss for the feature extractor.

3.5 Segmentation with BCDU-Net

Segmentation is necessary to obtain Ground truth saliency maps for large scale datasets without the expert knowledge of a dermatologist. Later on, saliency maps will be quantified in terms of their meaningfulness based on the overlap with the actual skin lesion. In this work image segmentation is performed using the BCDU-Net [27] with the aim of saliency quantification. BCDU-Net is based on U-Net [46], which was widely use in medical image segmentation [47]. Initially introduced in 2015, the U-Net architecture was designed specifically for the task of biomedical image segmentation. At its core, the U-Net architecture consists of three key components: an encoder, a bridge, and a decoder. The encoder component comprises a series of convolutional layers, responsible for iteratively reducing the spatial dimensions of the input image. The bridge serves to connect the encoder and decoder components. The decoder is comprised of deconvolutional layers which increase spatial resolution as well as reducing number of channels. The final output shares the same resolution as the input image, with each pixel being assigned a specific class label. Segmentation tasks require the model to capture both high-level features, such as identifying objects, and low-level features, like detecting edges. Due to the challenge of simultaneously capturing these features while reducing spatial dimensions, it becomes obligatory to propagate high-resolution spatial information from the initial layers to the deepest layers. This is where the U-Net's copy and crop mechanism becomes significant, as it directly introduces high resolution information to the deepest layers. This mechanism enables the achievement of high-resolution segmentation, allowing for the precise delineation of complete objects [46].

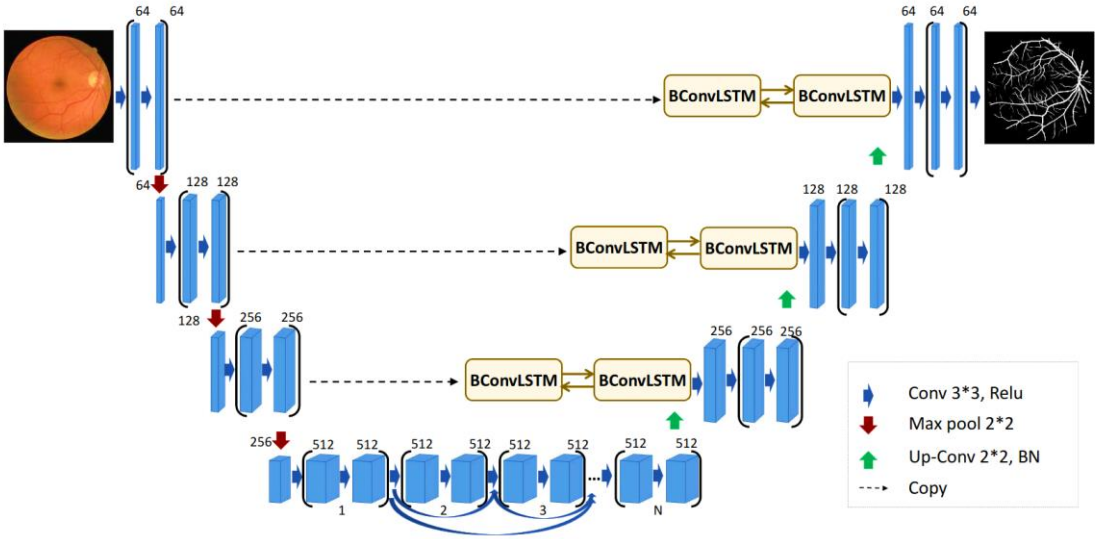

Figure 9: BCDU-Net architecture [27]

The Bi-directional Convolutional Long Short-Term Memory U-Net with Densely Connected Convolutions is a fusion of Convolutional Long Short-Term Memory (ConvLSTM) and U-Net architectures, incorporating densely connected convolutions within the bridge component of the U-Net framework. This architecture operates similarly to U-Net with the distinctive feature of the Bi-directional ConvLSTM layer intersecting the copy paths of the original U-Net (see Figure 9). By introducing these architectural augmentations, the BCDU-Net demonstrates enhanced segmentation precision alongside accelerated computation, primarily facilitated by the integration of batch normalization [27].

3.6 Statistical Metrics

3.6.1 Binary Classifier Metrics F1-Score, Specificity, Sensitivity

When characterizing the effectiveness of a binary classifier, in addition to accuracy, there are several other metrics that are of high significance.

Sensitivity, denoted as the probability for a positive sample being recognized as such and is formally defined as: $Sensitivity = \frac{a}{a+b}$ [48] (4)

a represents the number of true positive classifications, whereas b represents the number of false negative classifications.

Specificity, regarded as the probability for a negative sample being recognized as such and is formally defined as: $Specificity = \frac{d}{d+b}$ [48] (5)

Here, d stands for the count of false positive classifications.

The F1-Score measures the classification performance in cases of imbalanced test data.

It is formally defined as: $F_1 = \frac{2a}{2\,a+b+c}$ [48] (6)

3.6.2 Similarity Metric: Jaccard Index (IoU), wIoU

To quantify the similarity between two entities A and B, the Jaccard Index serves as an appropriate tool. The Jaccard Index is also known as Intersection over Union, which directly descibes the matematical essence of its concept.

$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$ [49] (7)

The weighted Intersection over Union (wIoU), employed in this work, is not be confused with the weighted Jaccard similarity, which is applicable to two vectors.

$wIoU(x,y) = \frac{\sum_i x_i * y_i}{\sum_i x_i + y_i - x_i * y_i} \neq J_w(x,y) = \frac{\sum_i \min(x_i,\ y_i)}{\sum_i \max(x_i,\ y_i)}$ [49] (8), (9)

With x, y representing vectors with i elements $\forall\, x_i, y_i \in \mathbb{R} \geq 0$

The wIoU can be regarded as the ratio of the cumulative weighted computation of overlap (numerator) and union (denominator) across all elements. This calculation considers the elementwise interaction of these weighted values.

Weighted IoU is necessary for adequately quantifying the output of a saliency method, since these values are not binary. Utilizing the conventional IoU metric in this context would introduce a new challenge of determining an appropriate threshold for binarization. However, incorporating the importance of each element within a saliency heatmap would produce results that closely correspond to the attention of the CNN.

4. Methods:

4.1 Classification of Skin Lesions with a Pretrained ResNet18

The classification aims at discerning melanoma and nevus images. This classification solely relies on the image data without using contextual patient information. The most notable system characteristics for the classification algorithm are Python 3.9, PyTorch 1.13.1 and ScikitLearn 1.2.1. If not separately specified, the Python and PyTorch dependencies were always utilized for all methods. The seed for reproducibility was 0 for PyTorch and NumPy.

In terms of architecture, a pretrained ResNet18 model was employed as a classifier and subjected to a training spanning 10 epochs. The final layer of this architecture consists of two output neurons for the binary classification task. The output tensor contains two float values, which is then passed to the torch.max function. This yields the index of the highest value, which represents the output neuron with the highest activation and therefore completes the classification process.

In this work, diverse datasets were applied. A big subset of the ISIC- Dataset , spanning over multiple domains [50] and later referred to as the inter-domain dataset, as well as domain-specific ISIC datasets, [51]–[53]. These datasets were utilized in various classification experiments as depicted in Figure 10. Notably, the primary classification of skin lesions was conducted with models that were trained on the same dataset as they had been tested on. Domain adaptation was investigated in two different ways. In the first approach, one source domain provided the training data for a model, which was subsequently tested on the test sets of all other ISIC domains. In this context, dataset augmentation and DANN were inspected for their domain adaptation capabilities. The second approach involved training a model for each available ISIC domain on the respective domain-specific training dataset. The resulting numerous models were tested on their corresponding testing dataset before their domain adaptation behavior was assessed on the Human Against Machine (HAM) loc body test set. Additionally, the impact of an augmented test dataset was investigated.

During preprocessing, all images underwent normalization. The reason behind the normalization is the faster convergence with better and stable results [54].

For training, the PyTorch cross-entropy loss function and an optimizer using Stochastic Gradient Descent (SGD) was applied.

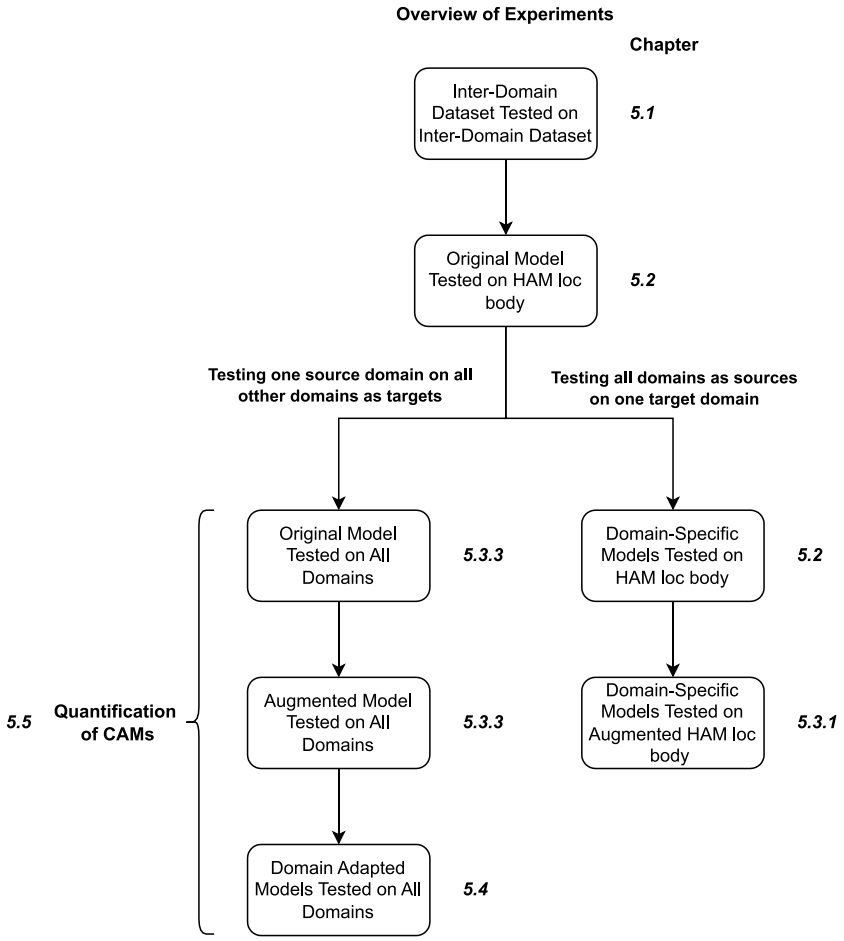

Figure 10: Overview of Classification Experiments (created with draw.io)

The classification process, depicted in Figure 11, commenced with the loading of ISIC image data. The ISIC-Dataset was loaded using the DataLoader function in PyTorch. Detailed information on accessing data within the domain structure will be provided in a subsequent subsection of this chapter. For the inter-domain ISIC-Dataset, image labels were available. The training dataset used for the presented results comprises 14,791 images, including 3,812 melanoma and 10,979 nevi samples. If training had already been conducted, a pre-trained model could be loaded with respective parameters. Otherwise, the training algorithm had to be applied as illustrated in Figure 11.

The training process involved iterating over the available training data, which is divided into batches, for a predefined number of epochs. During each iteration, batches were fed into the model, and the batch loss was calculated by comparing the model's classifications to the image labels. The training progress was surveilled by plotting this loss. Gradients were computed based on this loss, which guided the optimization of model parameters. Upon completion, the model parameters were saved. Loading the model for testing requires the exact same network architecture used during training.

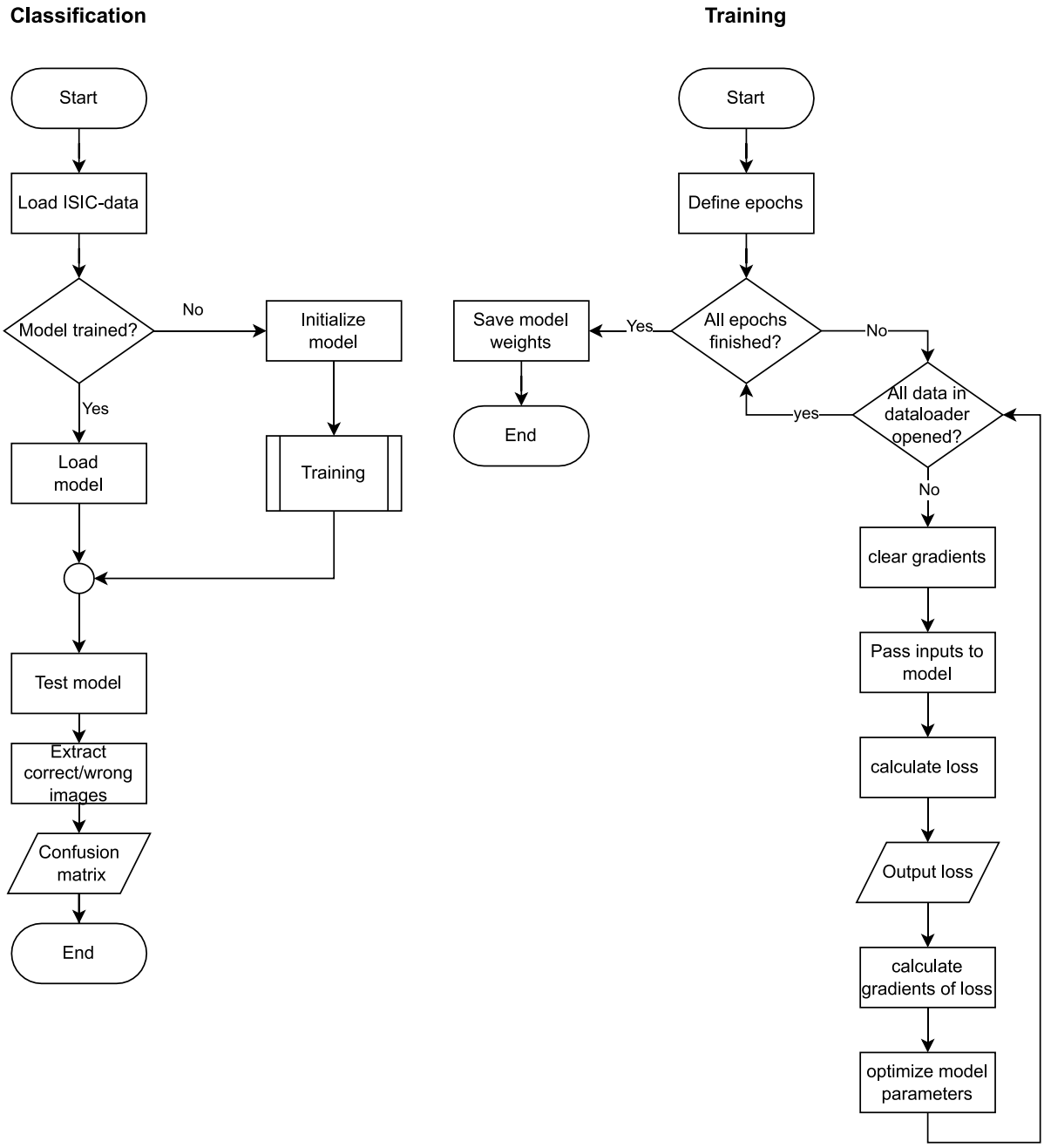

Figure 11: Flowchart of the employed training process (right) and the classification algorithm(left) (created with draw.io)

Subsequently, validating the model on the holdout test data of the training set was necessary to evaluate the actual performance of the model.

After the model had been examined for training success, testing the model on the desired dataset was used to exactly determine which images were classified correctly or incorrectly. From this information a confusion matrix could be derived.

4.2 Grad-CAM for Dermoscopic Datasets

To utilize Grad-CAM on specific images, several prerequisites had to be fulfilled. Firstly, a trained classification model needed to be chosen. Secondly, the input data, represented by an image along with its true classification label, needed to be available. Knowing the image's predicted class was crucial for identifying disparities between correct and incorrect image classifications. Hence, the correct and wrong classifications were saved in separate data subsets. This was essential for later analysis. The most notable system characteristics of this Grad-CAM implementation are Torchvision 0.14.1, NumPy 1.24.0, Pytroch-Grad-CAM 0.2.1 and OpenCV-Python 4.7.0.68.

The two subsets, previously generated during the classification process, were fed through the network. The resulting activations in the target layer of ResNet18 were utilized by Grad-CAM to create individual saliency maps for each image. An essential preprocessing step involved the denormalization of the images that were to be plotted. Otherwise the images would have occurred excessively dark and would not have been suited for human perception.

Grad-CAM continued to operate on the normalized version of the image. The target class, for which Grad-CAM activations were displayed, corresponds to the model's classification output. To visualize the results, the Matplotlib Python library was employed to create image matrices for each subset.

Up to 40 images were fed to the Grad-CAM algorithm for each subset to gain a comprehensive overview of the activation patterns. For quantification purposes, Grad-CAM was applied to the whole test dataset to obtain representative values.

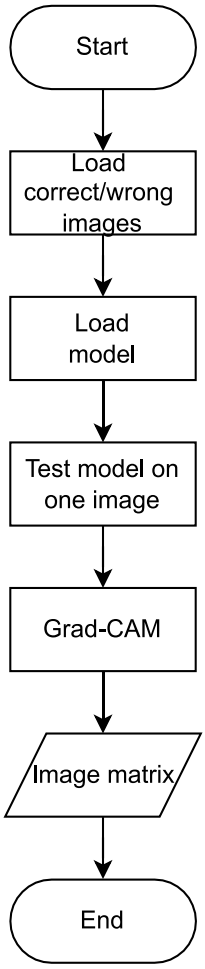

Figure 12: Flowchart of the Grad-CAM implementation (created with draw.io)

4.3 Accessing the Domain Data Structure

In this section, the implementation of accessing the domain data structure presented in Table 1, which was proposed by Fogelberg et al. [10], will be elaborated. The most notable system characteristics are Torchvision 0.14.1, NumPy 1.24.0 and ScikitLearn 1.2.1.

To access the data of one particular domain, it was essential to define the desired domain and its corresponding parameters. This was accomplished by reading a Comma-Separated Values (CSV)-file or manually specifying the parameters. The parameters included batch size, folder paths and file names. Subsequently the datasets were split into train and test sets. The source path was determined using the previously defined parameters along with the available metadata. From this source path the images were then loaded via the matplotlib imread function. At this stage, various image transformations were applied, depending on the experiment. Image normalization was always performed in this process. Finally, the PyTorch DataLoader was utilized which employed a weighted random sampler to obtain an equal number of images for both classes, while dividing them into batches. The random sampler's purpose was mitigating class imbalance and its application usually resulted in the repeated loading of melanoma images, since it is the underrepresented class in the majority of datasets. This data-loading procedure was applied in all experiments that are depicted in Figure 10.

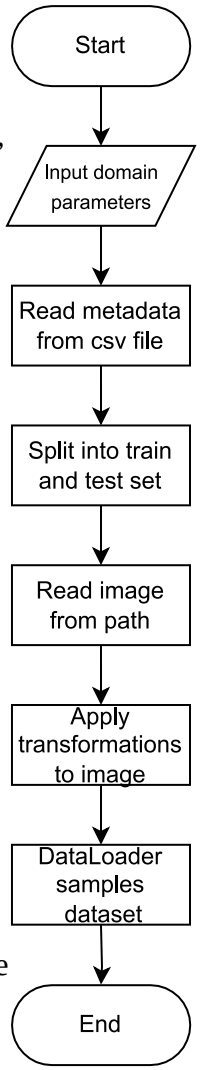

Figure 13: Flowchart of the domain access process (created with draw.io)

Table 1: Overview of the examined ISIC domains [10]

Origin	Biological factors	Melanoma amount	Nevus amount	Total target size
HAM	Age >30, Loc. = Body (default)	465 (10%)	4234 (90%)	4699
HAM	Age ≤ 30, Loc. = Body	25 (4%)	532 (96%)	557
HAM	Age >30, Loc. = Head/Neck	99 (45%)	121 (55%)	220
HAM	Age >30, Loc. = Palms/Soles	15 (7%)	203 (93%)	218
BCN	Age >30, Loc. = Body (default)	1918 (41%)	2721 (59%)	4639
BCN	Age ≤ 30, Loc. = Body	71 (8%)	808 (92%)	879
BCN	Age >30, Loc. = Head/Neck	612 (66%)	320 (34%)	932
BCN	Age >30, Loc. = Palms/Soles	192 (65%)	105 (35%)	297
MSK	Age >30, Loc. = Body (default)	565 (31%)	1282 (69%)	1847
MSK	Age ≤ 30, Loc. = Body	37 (8%)	427 (92%)	464
MSK	Age >30, Loc. = Head/Neck	175 (60%)	117 (40%)	292

4.4 Augmentation of Dermoscopic Images

As mentioned earlier, image augmentations significantly impact the resulting model's performance. Augmentation was applied to both the HAM loc body domain as the target domain and for HAM loc body as the source domain. Specifically, the target domain augmentation was achieved by solely augmenting the test set of the target domain, while the source domain augmentation involved augmentation of both the source domain's training set and the target domain's test set. The applied augmentations differed in those two scenarios. In the case of HAM loc body as the target dataset, the primary goal of augmentation was to eliminate a geometric bias inside the nevus class, which had been observed during the experiments. In this scenario, geometric augmentations were exclusively applied to nevus images using Torchvision transformation functions, because no bias has been discernible in melanoma samples. Specifically, a random crop of 315 pixels in height and 420 pixels in width was applied first [55]. While maintaining the original image's aspect ratio, this augmentation function randomly selected a 315x420-pixel area inside the skin lesion image. As a result, features or objects located in the image's peripheral areas were randomly excluded from the augmented image, forcing the model to focus on the skin lesion. Subsequently, the image was flipped horizontally or vertically with a respective probability of 0.5 for the execution of the transformation [56], [57]. The image was then resized to the original scale of 450x600 pixels using the bilinear interpolation with antialiasing enabled [58]. Consequently, objects or features within the image were relocated while maintaining the original image size.

Resizing to a lower resolution of 224x224 pixels and therefore altering the aspect ratio was an additional option to avert learned patterns that are associated with image size and aspect ratio, like frame-patterns. This however negatively affects performance since the model is confronted with lower resolution images although it was trained on higher resolution images. In tasks, which aim to maximize performance, resizing of the test images was not an option.

In the final step, the Torchvision normalize function was applied to both nevus and melanoma images. Normalization involved standardizing each color channel of a red, green and blue (RGB)-format image, represented by a float input tensor with values in the range [0;1], to fit the values in a range of [-1;1]. Since the Torchvision normalize function is defined as:

$$output = \frac{input - mean}{std} \qquad [59] \qquad (10.1)$$

With std representing the standard deviation, while mean and std being manually specified as 0.5.

The results for the maximum or minimum values were derived like this:

$$output_{max} = \frac{1-0.5}{0.5} = 1 \quad (10.2)$$

$$output_{min} = \frac{0-0.5}{0.5} = -1 \quad (10.3)$$

This transformation was essential since pretrained ResNet models rely on normalized values, distributed across a small negative and positive range. Otherwise, the transfer learning approach of ResNet would add limited benefit to the training process.

It is common practice to augment the underrepresented class within a dataset in combination with a weighted sampler to mitigate class imbalance. Therefore, additional augmentation was applied on HAM loc body as the source dataset, to address the underrepresentation of melanoma. In this scenario, all the previously mentioned Torchvision functions were also applied to the melanoma images, except for random cropping. As illustrated in Figure 18, melanoma images typically cover a larger area in the images, and cropping could result in the loss of crucial information. The images were resized to 224x224 pixels to precisely match the standard input size of ResNet18. Otherwise, an AdaptiveAvgPool Layer of ResNet18 adjusts the output size of the network's Layer4 to the desired ResNet18 size [60]. Resizing during preprocessing was viable since the augmented data was utilized to train a new model in this scenario.

4.5 Implementation of DANN

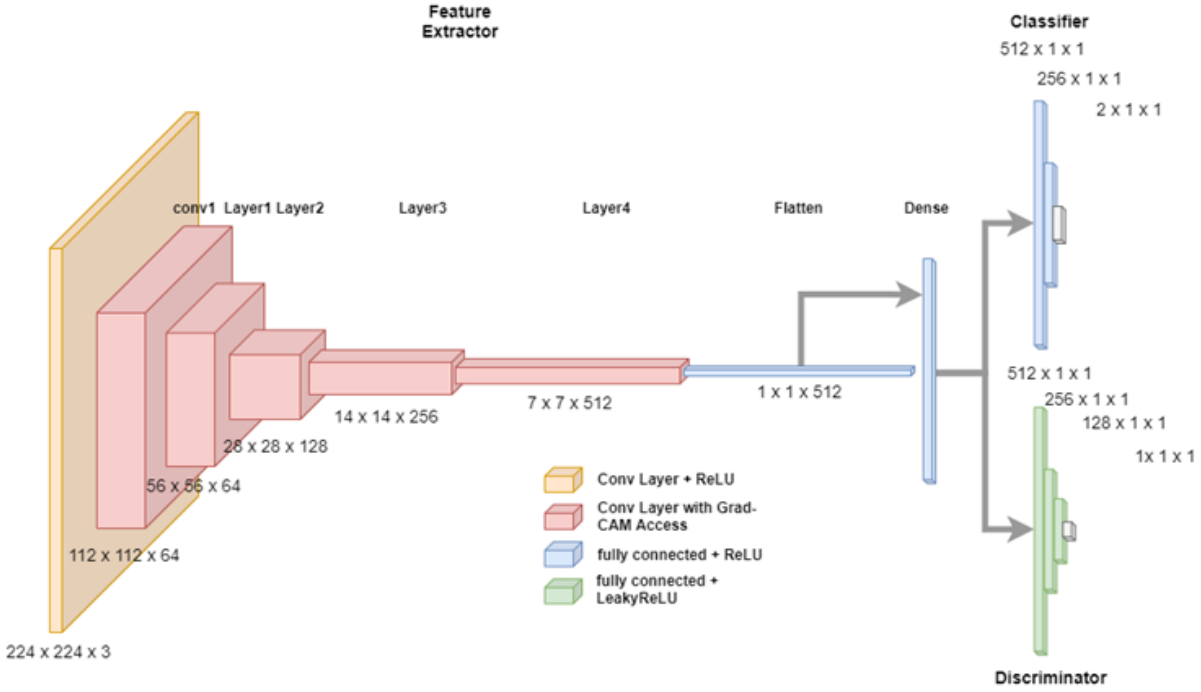

Figure 14: Layer dimension representation of the implemented DANN architecture with a ResNet18 Backbone (created with draw.io)

The implementation details of the previously described DANN principle are explained in this section.

The DANN architecture employs multiple networks for unsupervised domain adaptation. A feature extractor, referred to as the "backbone" network, is a crucial component. In this context, the ResNet18 has been chosen to maintain comparability to the augmentation approach (see Figure 14). Therefore, the DANN architecture specified in this DANN GitHub [3]implementation [61] was modified to process ISIC image data. The classifier was comprised of fully connected layers. Likewise, the Discriminator received the same ResNet18 output and utilized LeakyReLU activation functions. LeakyReLU is used to gauge negative activations instead of completely disregarding them. Ultimately, the Discriminator was utilized to distinguish between the domains via a sigmoid activation function at the output neuron. The classifier applied a softmax at the output layer, with a number of output neurons corresponding to the number of classes.

The Classification workflow was similar to the previously described classification with ResNet18. However, there were specific training process details, which were crucial for the DANN implementation, as illustrated in Figure 15. Each batch consisted of images from both domains, which were fed through the feature extractor, generating feature vectors. All vectors were then passed to the discriminator along with the corresponding domain label. Subsequently, backpropagation was performed through the discriminator, calculating its loss and the gradients. In parallel, only the feature vectors of the source domain were being passed to the classifier for the computation of its loss and gradients, since class labels were not provided for the target domain. A total loss was then calculated from the discriminator and classifier losses. Due to the predefined gradient reversal, an epoch-dependent factor lambda was used to gauge the subtraction of the discriminator loss from the classifier loss. With all losses and gradients available, parameter optimization was performed.

[3] https://github.com/Yangyangii/DANN-pytorch/blob/master/DANN.ipynb

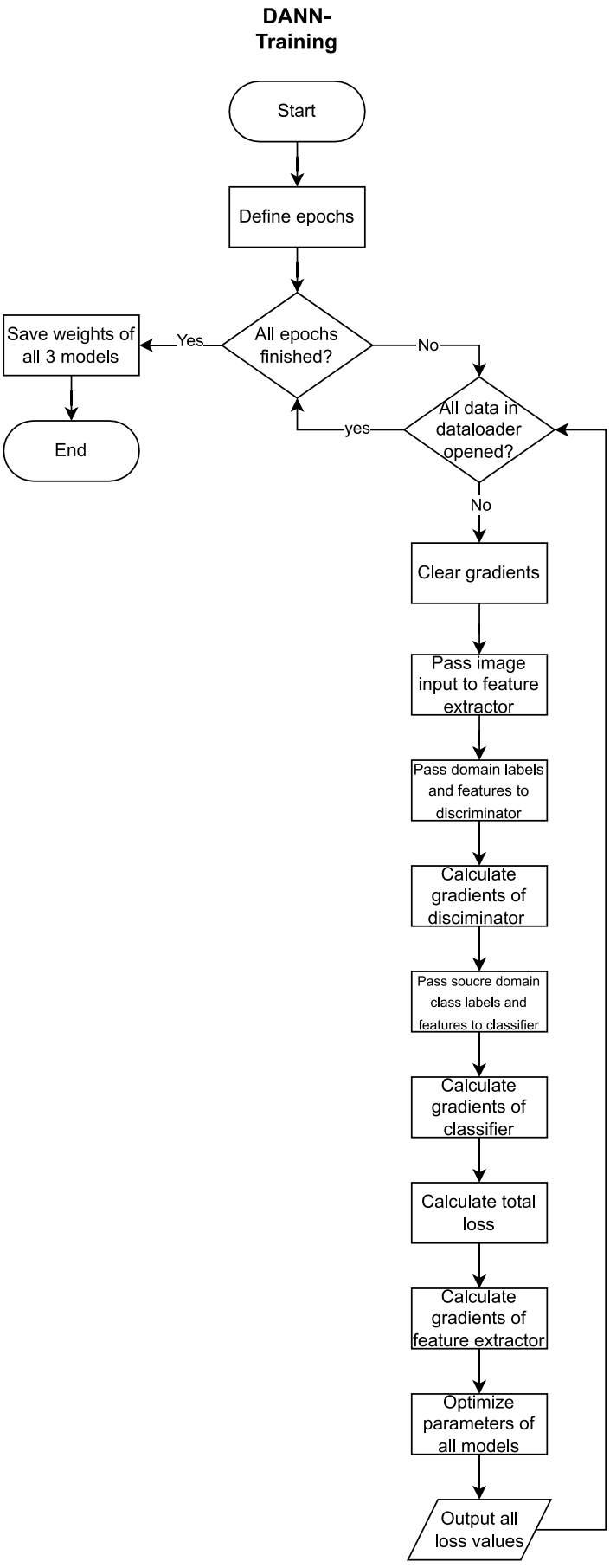

Figure 15: Flowchart of the DANN training process (created with draw.io)

4.6 Implementation of BCDU-Net

For the purpose of saliency quantification, the ISIC samples inside the domains had to be segmented first. Given that BCDU-Net therefore served as the foundation for quantification, an explanation of its implementation is necessary. The BCDU-Net was implemented following the GitHub [4]repository by Reza Azad [62]. Minor adjustments were made to address deprecated libraries. Notable system characteristics are Tensorflow 2.12.0, Keras 2.12.0, NumPy 1.23.5 and ScikitLearn 1.2.2.

The BCDU-Net implementation for skin lesions is comprised of three main phases, preparation, training and evaluation. During the preparation phase, the ISIC 2018 Challenge training data [8] was prepared for training. This dataset consists of dermoscopic images and their corresponding ground truth binary segmentation masks, created by professional dermatologists. First the images were read from the dataset using the imageio read function. After resizing, these images were systematically appended to a NumPy array. The same procedure was applied to the segmentation masks. Subsequently, the resulting NumPy data-frames were partitioned into training, validation and test sets. Upon completion of partitioning, the resulting six data frames were saved to the system.

During the training phase, the image data frames were normalized based on the mean and standard deviation computed from the NumPy data frame. Additionally, the dimensionality of the mask data was expanded to align with the RGB format of the image data. The BCDU-Net was loaded and trained to fit the image data to the masks. Throughout the training process, only the model with the lowest validation loss was saved.

In the evaluation phase, the model predicted masks based on the image data input. These predicted masks, initially in RGB format, were subsequently reformatted into binary masks. The conversion to binary masks facilitates the computation of informative metrics for assessing performance, such as the Jaccard score, F1 score, and the generation of a confusion matrix. In addition, sample images with ground truth masks along with their predictions can be displayed to visually evaluate the model's performance.

In this work, the objective was not to predict ground truth data. Instead, the developed model served the purpose of predicting a specific segment within a skin lesion to quantify the plausibility of saliency. The workflow can be divided into a preparation and evaluation phase.

During the preparation phase, the ISIC image data had to match the structure of the training data for BCDU-Net. To achieve this, the ISIC images from a specific domain were resized to

[4] https://github.com/rezazad68/BCDU-Net/tree/master

224x224 pixels and were organized into a NumPy data frame. Subsequently, Grad-CAM was applied to all images in the dataset, generating class activation maps, which were resized to 224x224 pixels and segregated into two NumPy data frames, depending on correct or wrong prediction.

To determine the actual performance of BCDU-Net, the computation method of the Jaccard score has been altered. Since the measured value was derived from the entirety of the images, larger skin lesions had a greater influence on the result. Computing the IoU for each image and averaging over all IoU values yielded a more representative result. Therefore, the method was modified for better validity.

The evaluation algorithm has been altered to measure the alignment between the saliency predictions and the skin lesion segment. At this stage, images from one domain were fed through the model. The binary segmentation predictions generated by BCDU-Net were compared to the previously saved class activation maps instead of ground truth segmentation masks. This entails one fundamental difference to the predefined evaluation procedure. The class activation maps are greyscale images that contain floating-point values between 0 and 1. A higher value represents greater activation in the corresponding image region. For reasons detailed in the 3.6.2 Similarity Metric: Jaccard Index (IoU), wIoU subsection, binarization of the class activation maps is not a viable option. Since F1-Score and confusion matrix are bound to a binary classification problem, they cannot add benefit to the quantification of class activation maps plausibility. To address this challenge, the Jaccard score has been adapted to take float values into account. This resulted in a modification referred to as “wIoU”. The specifics of this modification are elaborated in the subsequent section.

4.7 Implementation of wIoU

The intensity of the CAM activations is taken into consideration by the wIoU, high activation within the skin lesion segment strongly contributes to the wIoU value of a sample. The wIoU served the purpose of averting the threshold optimization problem, as discussed in section 3.6.2. This optimization would have to be applied to every domain, limiting the comparability of the measured values in case of different thresholds. In the context of a single image with its corresponding segmentation masks and class activation map, the intersection and union values were computed. Intersection was derived from the element-wise multiplication of all mask values (binary) with the corresponding class activation map values (float). Union was calculated by summing mask and activation map values while subtracting the intersection value. Intersection was subsequently divided by union, yielding a result that is in conformity with the

mathematical description in 3.6.2. Specifically, a wIoU of 1 is reached in case of exclusive maximum activation overlapping perfectly with the skin lesion segment However, due to the characteristics of the ResNet CAMs, it was unlikely to obtain a wIoU score of 1. This should be considered when interpreting the wIoU results. As previously mentioned, the wIoU has been calculated individually for each image and was then aggregated into a list. Subsequently, the mean of all wIoU values was determined from this list. This process yielded separate mean wIoU values for correct and incorrect classifications.

5. Results

5.1 Classification and Activations on Dermoscopic Datasets

Primary classification has been performed using the ResNet18, which was trained for 10 epochs on an inter-domain ISIC dataset, comprised of 14791 training images. The only image transformation applied before training is resizing the images to 224x224 pixels. The statistical results obtained from this evaluation are presented in Table 2.

Table 2: Test results on an inter-domain test dataset, containing 2385 images. Columns from left to right: accuracy, melanoma true positive, melanoma false negative, melanoma true negative, melanoma false positive, percentage of melanoma in training data, F1-Score for melanoma class.

Dataset	Acc.	Mel. tp	Mel. fn	Mel. tn	Mel. fp	Mel. %	F1
Inter-domain ISIC	0.885	0.7	0.3	0.954	0.046	0.551	0.768

The classifiers accuracy and F1 score in Table 2 show how a ResNet18 classifier performs with access to multiple domains. Despite a slight overrepresentation of melanoma samples in the training dataset, the resulting classifier was clearly worse at correctly identifying melanoma than it was at identifying nevus.

For the classifiers Grad-CAM heatmaps, a brief excerpt is presented in Figure 16 to outline the discovered patterns. In the case of correct melanoma classification, it is observed that the activation maps often do not precisely align with the skin lesion area. Neural network activation overlaps with the skin lesion area more accurately in case of correctly classified nevus samples. Neural network activation of wrongly classified actual nevus images is clearly less plausible than in cases of correct nevus classification. For incorrectly classified actual melanoma samples however, the activation plausibility is similar to correctly classified melanoma samples, since the activation area continues to not match the area of the skin lesion.

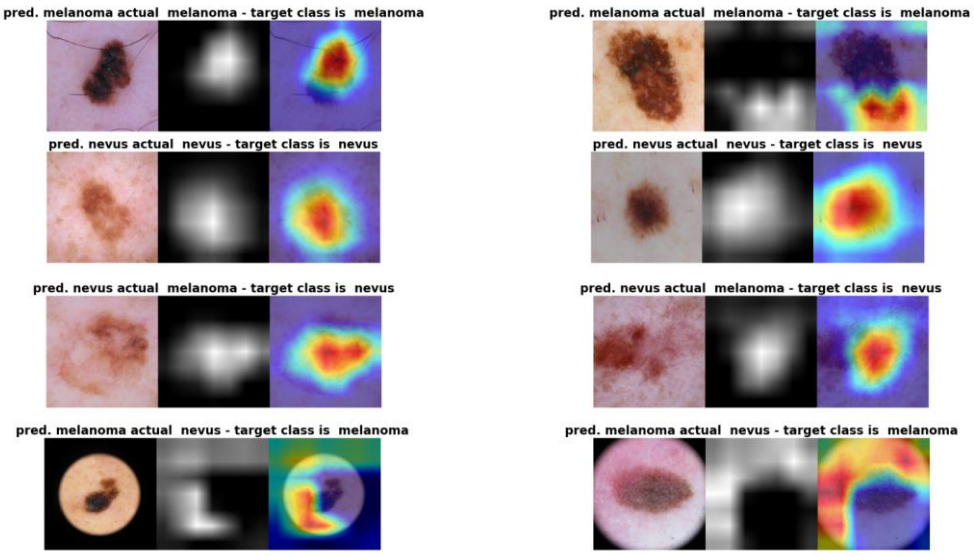

Figure 16: Representative samples of the inter-domain test dataset with their class activation map (middle image) and their heatmap overlay (right image). Top row: correctly classified melanoma, second row: correctly classified nevus, third row: falsely classified actual melanoma, fourth row: falsely classified actual nevus. (created with Matplotlib and GIMP 2.10)

5.2 Classification and Activations on Domain Shifted Datasets

To gain insights into the causes and manifestations of domain shift on image level, ResNet18 models underwent training on every available domain. Subsequently, these models were tested on both the source domain and HAM loc body as the target domain. Each model was exclusively trained for 10 epochs on its respective source domain training dataset.

Table 3: Test Result on HAM loc body as the source domain. Columns from left to right: melanoma share of training dataset in percent, accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class.

Dataset	Domain	Mel. %	Accuracy	Sensitivity	Specificity	F1
HAM	loc body	9.9	0.916	0.796	0.929	0.652

Table 4: Test Results on the source domains. Columns from left to right: melanoma share of training dataset in percent, accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class

Dataset	Domain	Mel. %	Accuracy	Sensitivity	Specificity	F1
HAM	age under 30	4.5	0.982	0.8	0.991	0.8
HAM	loc head neck	45.0	1	1	1	0.976
HAM	loc palms soles	6.9	1	1	1	1
BCN	loc body	41.3	0.971	0.964	0.982	0.97
BCN	age under 30	8.1	0.983	0.786	1	0.889
BCN	loc head neck	65.7	0.984	0.984	0.984	0.988
BCN	loc palms soles	64.6	0.967	0.974	0.952	0.974
MSK	loc body	44.1	0.965	0.938	0.977	0.942
MSK	age under 30	8.0	1	1	1	1
MSK	loc head neck	59.9	0.898	0.971	0.792	0.919

Table 5: Test Results on HAM loc body as the target domain. Every row represents a separate model, trained on the specified source domain. Columns from left to right: melanoma share of training dataset in percent, accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class

Dataset	Domain	Mel. %	Accuracy	Sensitivity	Specificity	F1
HAM	age under 30	4.5	0.904	0.118	0.991	0.196
HAM	loc head neck	45.0	0.917	0.774	0.933	**0.649**
HAM	loc palms soles	6.9	0.915	0.194	0.994	**0.311**
BCN	loc body	41.3	0.428	0.624	0.406	0.177
BCN	age under 30	8.1	0.905	0.065	0.998	0.119
BCN	loc head neck	65.7	0.909	0.667	0.935	**0.59**
BCN	loc palms soles	64.6	0.283	0.882	0.217	0.196
MSK	loc body	44.1	0.814	0.71	0.825	**0.43**
MSK	age under 30	8.0	0.893	0.022	0.988	0.038
MSK	loc head neck	59.9	0.627	0.839	0.603	0.308

The results in Table 3 reveal that the F1-Score of the ResNet18 model, that was trained on the HAM loc body dataset, is the lowest among all models tested on their respective source domains. It is even lower than the F1-Score of the Big ISIC model.

As shown in Table 4, all other domain specific models outperformed the inter-domain model when tested on their respective source domain test sets. Table 5 provides further insights regarding the evaluation of the retrieved models on the HAM loc body test set. Notably, five domains exhibited F1-Scores surpassing 0.3, with two of them achieving scores exceeding 0.5. Sensitivity has a significant influence on the F1-Score. Consequently, models that achieved greater sensitivity for melanoma on the HAM loc body test dataset also demonstrated superior F1-Scores.

The initial focus for the examination of the activation patterns will be on the on the HAM loc body test set, which also serves as the target domain for other models. By starting with the analysis of this domain, a foundation for understanding the challenges associated with image classification in this specific domain is established.

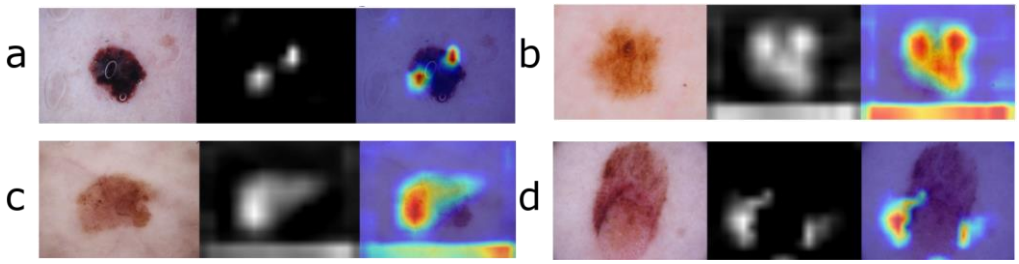

Figure 17: Representative samples of the HAM loc body test set with their class activation map (middle image) and their heatmap overlay (right image). a) correctly classified melanoma, b) correctly classified nevus, c) actual melanoma classified as nevus d) actual nevus classified as melanoma (created with Matplotlib and GIMP 2.10)

The class activation maps of correctly classified melanoma samples exhibit spot-wise activation and do not cover a substantial area of the skin lesion (see Figure 17). In contrast, CAMs for the nevus class often highlight meaningful regions within the image. Notably, there is a ubiquitous activation area present at the bottom of all accurately classified nevus images, frequently taking on a bar-shaped pattern.

In cases where actual nevus images were misclassified, the corresponding activation maps once again display spot-like activations, missing rational patterns. For wrongly classified actual melanoma images, CAMs exhibit similar patterns to activation maps of correctly classified nevus samples.

Shifting the focus away from the details of activation patterns, it appears that more pictures of the correctly classified images have a red hue, which is primarily associated with nevus samples. A detailed random examination of 50 nevus samples inside the HAM loc body training set revealed that 35 out of 50 exhibited an intense red tone. In contrast, a random sampling of 50 melanoma images inside the same dataset revealed only 5 images with a discernible red hue. Illustrative examples of the samples are provided in Figure 18.

Along with the red hue, another peculiarity becomes apparent within the HAM loc body training dataset. Melanoma skin lesions tendentially occupy a larger portion of the image. Specifically, the aforementioned bottom region of the image is homogenous and clear of skin lesions or other objects in 13 out of the 50 randomly selected melanoma samples. With 28 out of 50 samples the nevus images have more than double the relative frequency of clear and homogenous bottom portions.

Illustrative samples were selected from the source and target domain to showcase the CAM patterns established by the domain-specific models. The insights were drawn from a comprehensive analysis of up to 80 images per domain, allowing the identification of consistent activation patterns.

The CAMs of the models, with outstanding performance will be closer examined in the following paragraphs. CAMs of the model trained on BCN loc body will be elaborated as well.

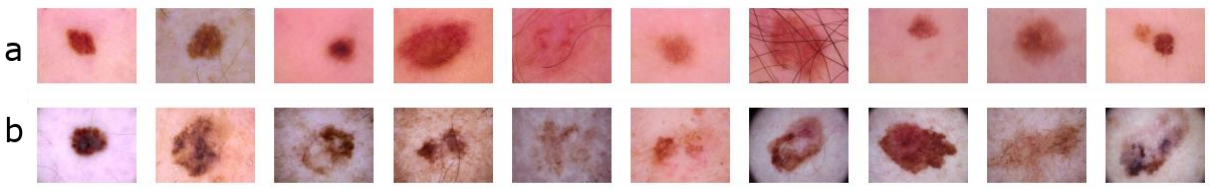

Figure 18: Random samples of a) nevus images and b) melanoma images of the HAM loc body training dataset. (created with Matplotlib and GIMP 2.10)

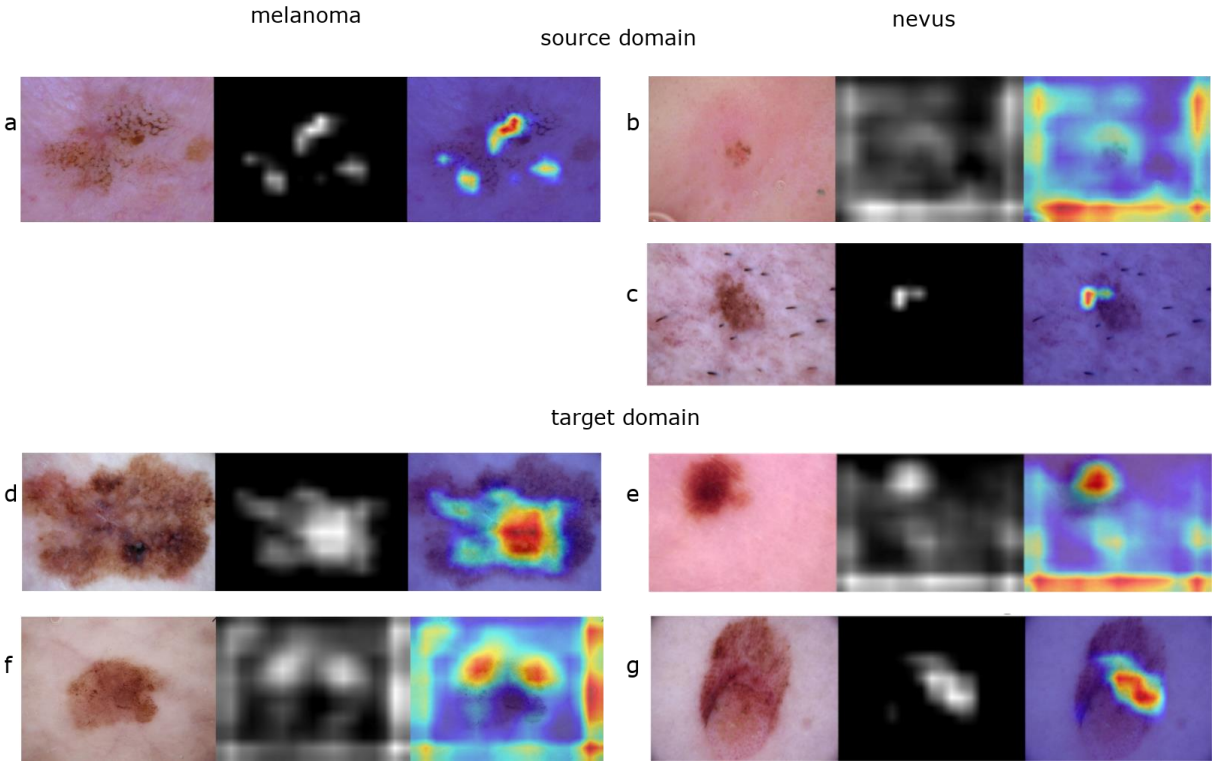

Figure 19: Representative samples of the HAM loc head neck test set with their class activation maps (middle image) and their heatmap overlay (right image). a) correctly classified melanoma, b) correctly classified nevus, c) falsely classified actual nevus is the only incorrectly classified sample. The lower half contains samples of HAM loc body test set, which is the target domain. d) correctly classified melanoma, e) correctly classified nevus, f) falsely classified actual melanoma g) falsely classified actual nevus (created with Matplotlib and GIMP 2.10)

The activations of the model trained on the HAM loc head neck domains training set are of great interest, since it exhibited the greatest F1-Score of all domain specific models tested on the HAM loc body test set. The examination of the source domain CAMs yielded that activation in melanoma samples is spot-wise but reasonably located on the skin lesion area (see Figure 19). CAMs of correctly classified nevus samples show a frame-like activation pattern in the periphery of the samples. Notably, the bottom-bar shape is also often present in the CAMs. The CAM of the only incorrectly classified source domain sample shows spot-wise activation on the skin lesion.

The CAMs on the target domain reveal activation with a focus on the skin lesion whilst not covering the complete skin lesion area in case of correct melanoma classification. For correctly classified nevus samples, the CAMs consistently exhibit a frame-like activation pattern, accentuated by a prominent bottom-bar. The remaining activation is often times meaningfully located on the skin lesion area. Wrongly classified actual nevus samples typically display spot-wise activation on the nevus images, which are sometimes not meaningful. CAMs of wrongly classified actual melanoma samples exhibit the mentioned frame pattern, although with less intensity compared to cases of correct classification.

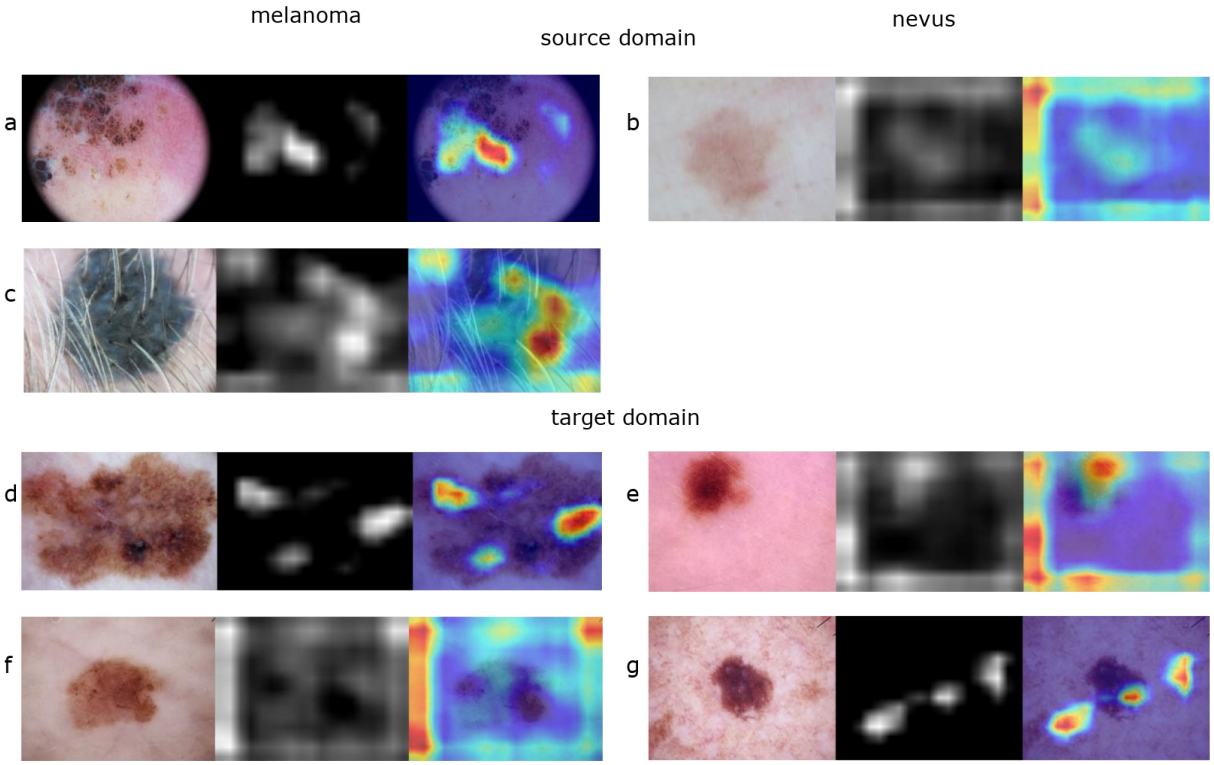

Figure 20:Representative samples of the BCN loc head neck test set with their class activation maps (middle image) and their heatmap overlay (right image). a) correctly classified melanoma, b) correctly classified nevus, c) the only available misclassification is an actual melanoma falsely classified as a nevus. The lower half contains samples of HAM loc body test set, which is the target domain. d) correctly classified melanoma, e) correctly classified nevus, f) falsely classified actual melanoma g) falsely classified actual nevus (created with Matplotlib and GIMP 2.10)

The second best F1-Score is achieved by the model trained on the BCN loc head neck training set. In this context, source domain CAMs for correctly classified melanoma samples display spot-wise activations in meaningful regions (see Figure 20). On the other hand, CAMs for correctly classified nevus samples lack clear purpose and often concentrate activation in the brighter peripheral areas of the images, resulting in a prominent frame-like shape present in most samples. The only misclassified source domain image contains a high amount of body hair. However, the activation appears more reasonable compared to many other samples for which the network decided nevus.

Examining the same model's CAMs on the target domain reveals that activations of correctly classified melanoma samples are more meaningful than in the source domain with larger activation spots on the skin lesion area. CAMs of correctly classified nevus samples always contain a frame pattern in which the left and bottom portions are especially prominent. In cases of wrong classification on the target domain, the class specific activation patterns persist, only differing in classification outcome.

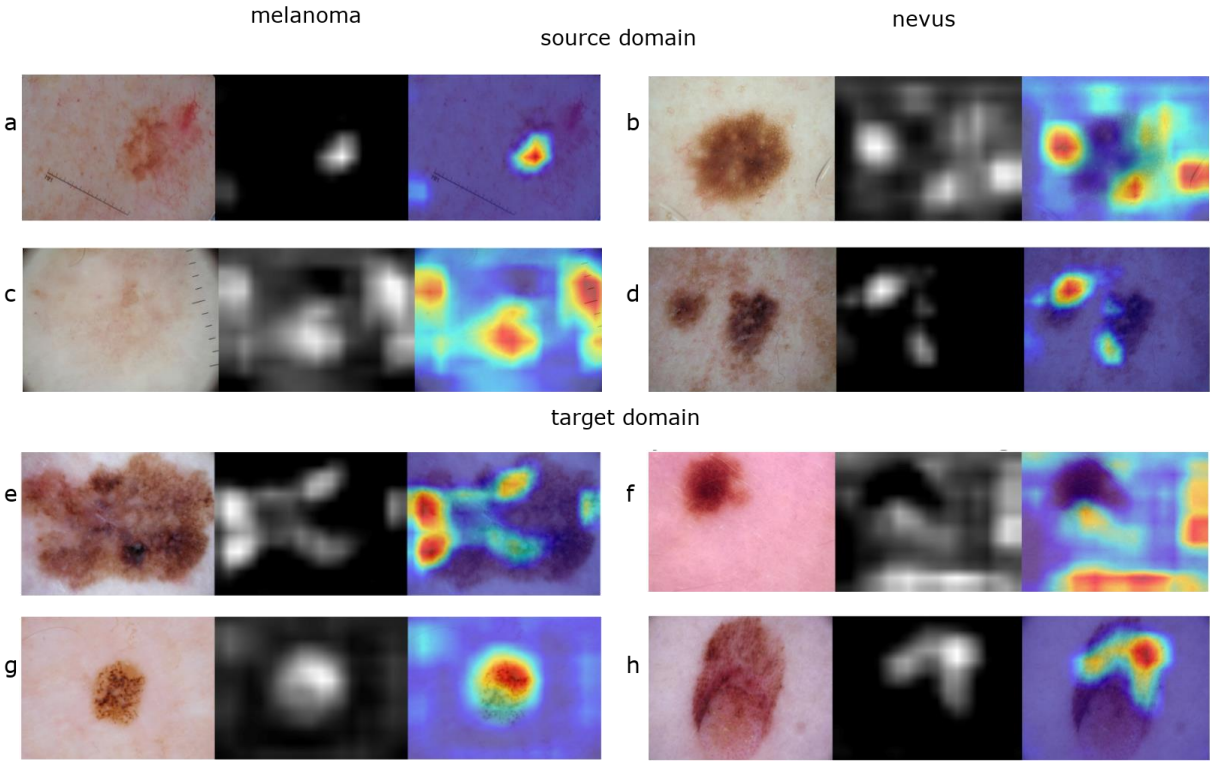

Figure 21: Representative samples of the MSK loc body test set with their class activation maps (middle image) and their heatmap overlay (right image). a) correctly classified melanoma, b) correctly classified nevus, c) falsely classified actual melanoma, d) falsely classified actual nevus. The lower half contains samples of HAM loc body test set, which is the target domain. e) correctly classified melanoma, f) correctly classified nevus, g) falsely classified actual melanoma h) falsely classified actual nevus (created with Matplotlib and GIMP 2.10)

The CAM examination of the MSK loc body model reveals that a high F1-Score and similar patterns are not restricted to the loc head neck domains (see Figure 21).

Correctly classified source domain melanoma samples contain meaningful activation but does not consistently cover the complete area of the skin lesion. CAMs of correctly classified nevus samples lack plausibility, since the CAM is often highlighting objects like rulers, bubbles or even bright areas in the peripheral regions. Plausibility of CAMs further diminishes in both cases of misclassification.

When the model, trained on the MSK loc body training set, is tested on the HAM loc body test set, the CAM patterns of the correctly classified melanoma don’t change. In contrast, CAMs of correctly classified nevus samples always contain a frame-like activation pattern that gains intensity when the activation area does not overlap with the skin lesion. Notably, the model did not exhibit this pattern on source domain samples. In case of wrongly classified actual nevus samples, the activation area usually has an overlap with skin lesions but does not cover large areas. CAMs of wrongly classified actual melanoma samples show plausible activation, despite the model's evident misclassification.

Given that HAM loc palms soles and MSK loc head neck also achieved an elevated F1-Score, it is reasonable to note discovered patterns in the CAMs of the test and on the target domain. Both models exhibit a frame-shaped activation pattern in case of nevus classification. This is especially true for the activations derived from testing on the target domain.

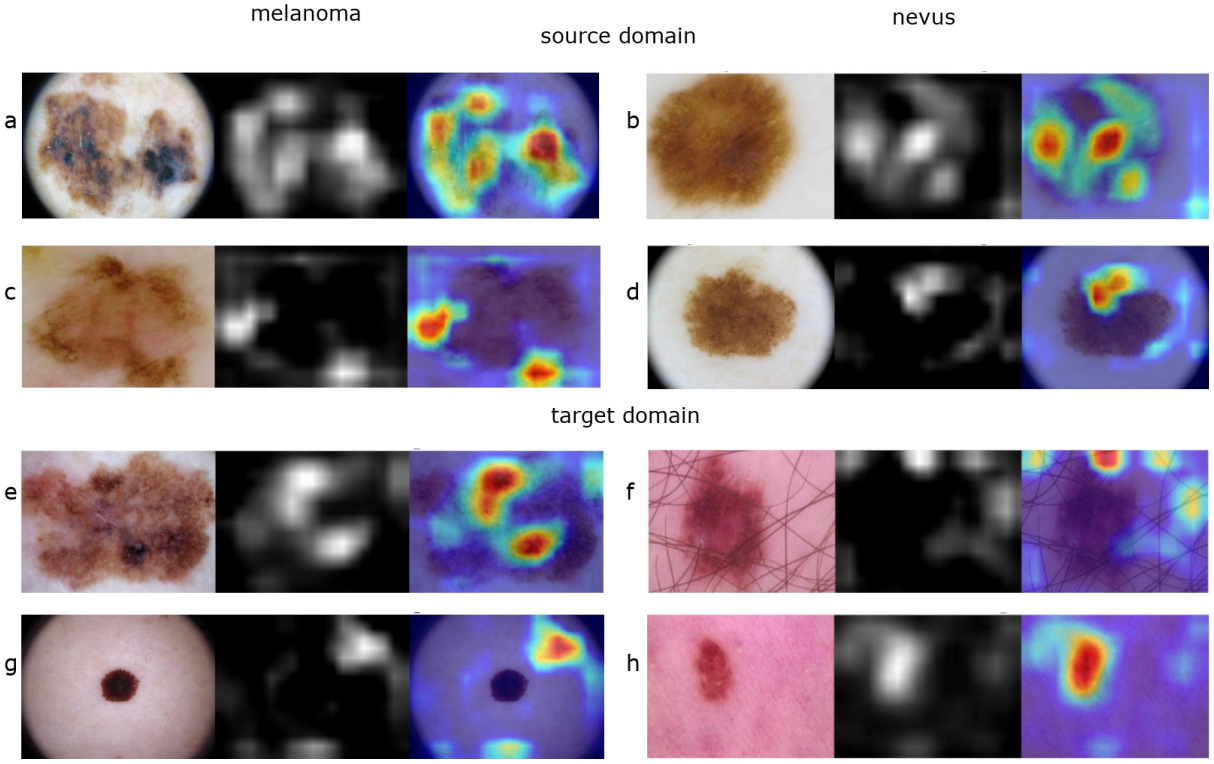

Figure 22: Representative samples of the BCN loc body test set with their class activation maps (middle image) and their heatmap overlay (right image). a) correctly classified melanoma, b) correctly classified nevus, c) falsely classified actual melanoma, d) falsely classified actual nevus. The lower half contains samples of HAM loc body test set, which is the target domain. e) correctly classified melanoma, f) correctly classified nevus, g) falsely classified actual melanoma h) falsely classified actual nevus (created with Matplotlib and GIMP 2.10)

Lastly, the CAMs of the BCN loc body model are investigated to find indications for the subpar performance on the target domain. The BCN loc body model indeed yields meaningful CAMs on its own test set, particularly for correctly classified melanoma samples (see Figure 22). In such instances, activations often cover larger areas of the skin lesion. Conversely, CAMs for correctly classified nevus samples exhibit more extensive but less intensive coverage compared to the melanoma CAMs. The wrong classifications tendentially show less overlap between the skin lesion and the activation area.

When the BCN loc body model is applied to the HAM loc body test set, the activation in correctly classified nevus samples is more easily diverted by other objects like body hair. In contrast, CAMs of correctly classified melanoma remain reasonable and activation covers substantial areas of the skin lesion. Notably, these patterns extend to the misclassifications,

manifesting in plausible activation in case of faulty decision for melanoma and distraction in case of faulty decision for nevus.

Shifting the scope away from the activations, a distinctive color mismatch between correctly and incorrectly classified samples inside the HAM loc body test set is observed. The aforementioned red hue inside the HAM loc body dataset is found to be overrepresented in the false classifications of the BCN loc body model.

5.3 Classification and Activations on Augmented Domain Shifted Datasets

5.3.1 Augmenting the Target Domain Dataset

Augmenting the HAM loc body test dataset aims at verifying the patterns that the detailed examination of the results in the previous section has yielded. With augmenting the nevus images inside the target domain without resizing the following results were derived.

Table 6: Test Results of the model trained on their source domain and tested on the augmented HAM loc body dataset without resizing. Bold F1-Score values indicate improvement compared to the non-augmented HAM loc body test set. Columns from left to right: accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class.

Dataset	Domain	Accuracy	Sensitivity	Specificity	F1
HAM	age under 30	0.913	0.118	1	**0.212**
HAM	loc head neck	0.972	0.774	0.994	**0.847**
HAM	loc palms soles	0.919	0.194	0.999	**0.321**
BCN	loc body	0.362	0.624	0.333	0.162
BCN	age under 30	0.906	0.065	0.999	0.12
BCN	loc head neck	0.901	0.667	0.927	0.571
BCN	loc palms soles	0.422	0.882	0.372	**0.232**
MSK	loc body	0.71	0.71	0.71	0.326
MSK	age under 30	0.889	0.022	0.985	0.037
MSK	loc head neck	0.652	0.839	0.632	**0.323**

From Table 6 it becomes evident that augmenting the HAM loc body test set cannot significantly improve overall classification performance compared to the non-augmented test set. Some already high performing models managed to enhance their performance on the augmented test set. A further examination of the Grad-CAM results is presented in Figure 23 with exemplary samples. Figure 23 elucidates the presence of the aforementioned frame pattern by showcasing an overview of correctly classified samples. Figure 23 further establishes an association between performance on the non-augmented HAM loc body test set to the intensity of the frame pattern in the CAMs of the augmented HAM loc body test set.

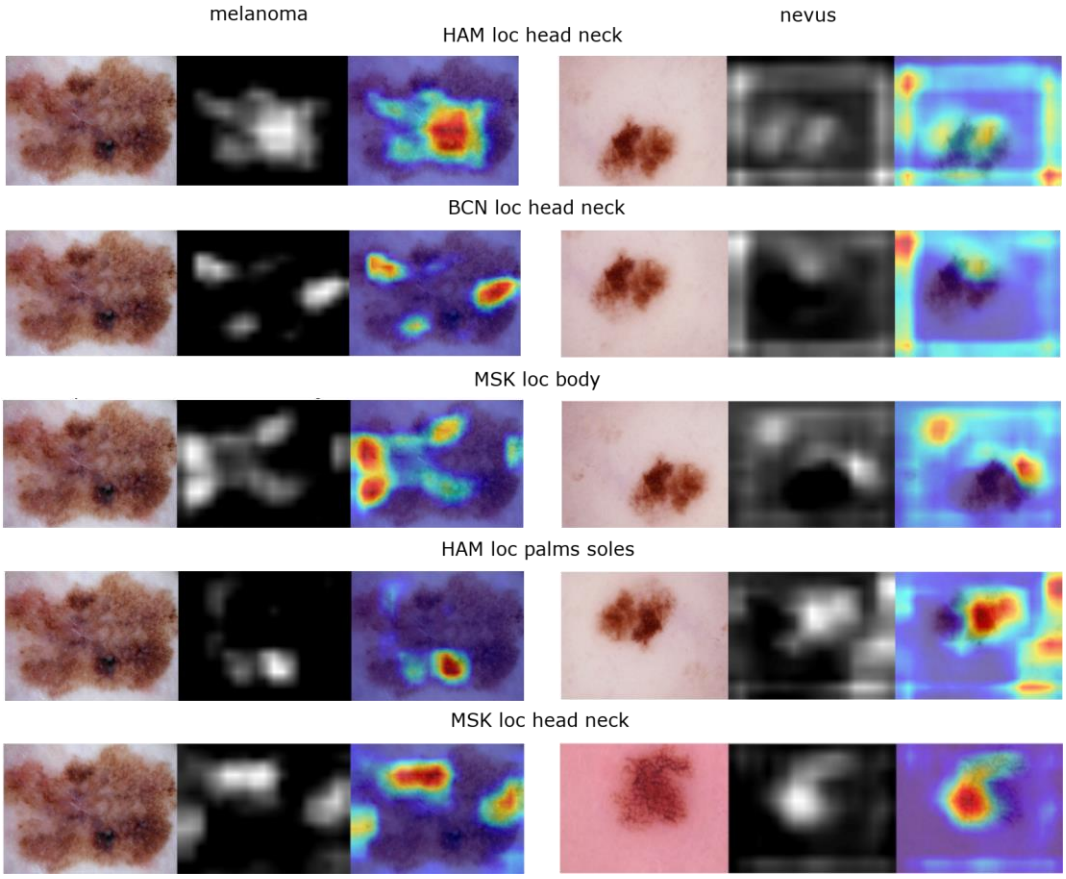

Figure 23: Representative samples from the augmented (no resizing) HAM loc body dataset with their class activation maps (middle image) and their heatmap overlay (right image). Employed domain-specific models in descending order of performance on the non-augmented HAM loc body test set. Left side melanoma, right side nevus. All presented images are correctly classified. (created with Matplotlib and GIMP 2.10)

Introducing resizing to 224x224 pixels to the augmentations the images exhibited the results in Table 7 and activations in Figure 24.

Table 7: Test Results of models, trained on their source domain and tested on the augmented (with resizing) HAM loc body dataset. Bold source datasets indicate high performance of according model on non-augmented HAM loc body test set. Bold F1-Score values indicate improvement compared to the non-augmented HAM loc body test set. Columns from left to right: accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class, absolute change of F1-Score compared to the test on the non-augmented HAM loc body test set.

Dataset	Domain	Accuracy	Sensitivity	Specificity	F1	F1 Change
HAM	age under 30	0.884	0.065	0.974	0.099	-0.097
HAM	loc head neck	0.701	0.935	0.675	0.382	-0.266
HAM	loc palms soles	0.916	0.591	0.952	**0.582**	0.271
BCN	loc body	0.395	0.71	0.36	**0.188**	0.011
BCN	age under 30	0.895	0.333	0.956	**0.385**	0.266
BCN	loc head neck	0.751	0.871	0.738	0.409	-0.181
BCN	loc palms soles	0.179	0.989	0.09	0.192	-0.003
MSK	loc body	0.44	0.753	0.406	0.21	-0.220
MSK	age under 30	0.879	0.022	0.973	0.034	-0.004
MSK	loc head neck	0.383	0.806	0.336	0.205	-0.102

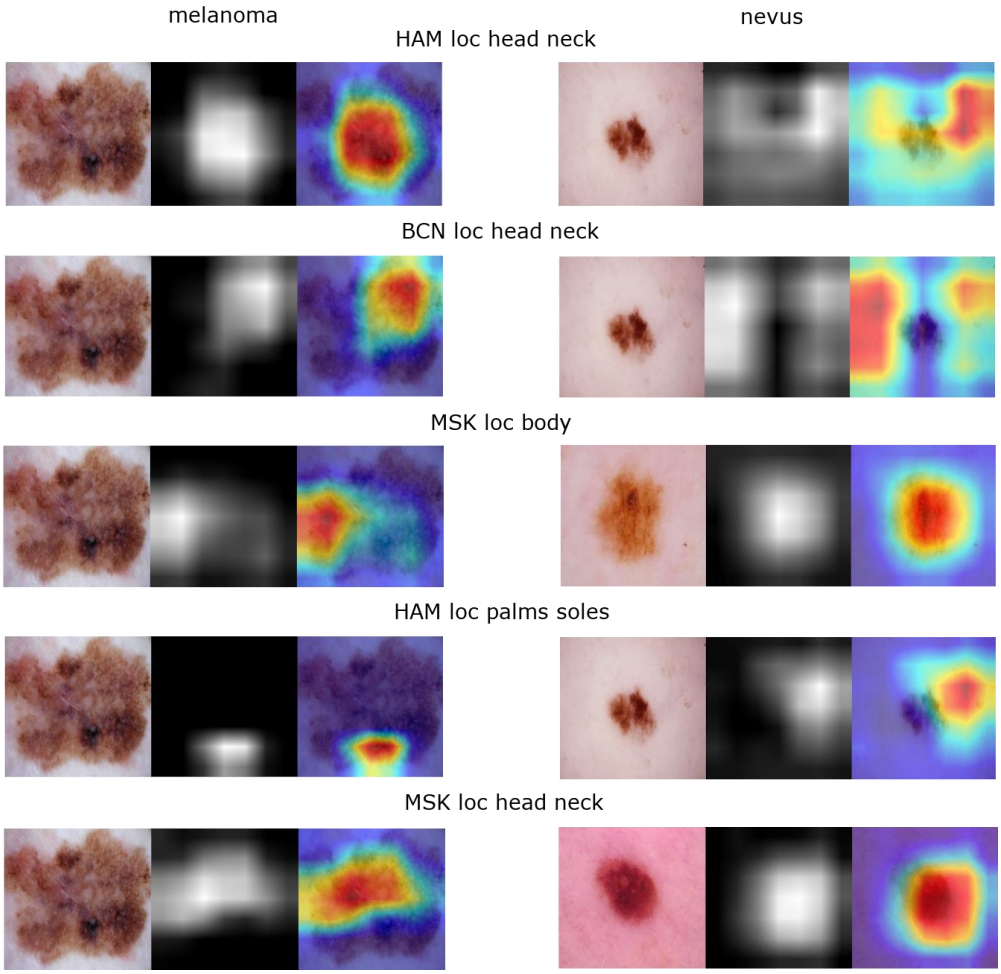

Figure 24: Representative samples from the augmented HAM loc body dataset with resizing, along with their class activation maps (middle image) and their heatmap overlay (right image). Every line shows CAMs for one model, trained on the specified domain. Left side melanoma, right side nevus. All presented images are correctly classified. (created with Matplotlib and GIMP 2.10)

The prominent frame shape pattern is not present in the resized images in Figure 24 anymore. The exact same nevus samples as in Figure 23 could not be provided since the samples were incorrectly classified by some models, and exhibit CAMs for the other class. The CAMs appear noticeably different to the ones derived from testing on the augmented HAM loc body test set without resizing. Activations are generally more plausible, while the overall statistical performance declines. It becomes evident that four out of 5 models which produce the frame shape activation pattern for nevus classification suffered a substantial reduction in their F1-Score when confronted with resized images. Conversely, other domain-specific models did not exhibit a similar decline in performance under these conditions. Notably, the model, trained on the HAM loc palms soles training set, demonstrated a significant F1-Score increase of 0.271.

5.3.2 Comparison of Saliency Methods

After initial Grad-CAM results of HAM loc body as the source domain with all other domains as the target domain the images in Figure 25 were obtained.

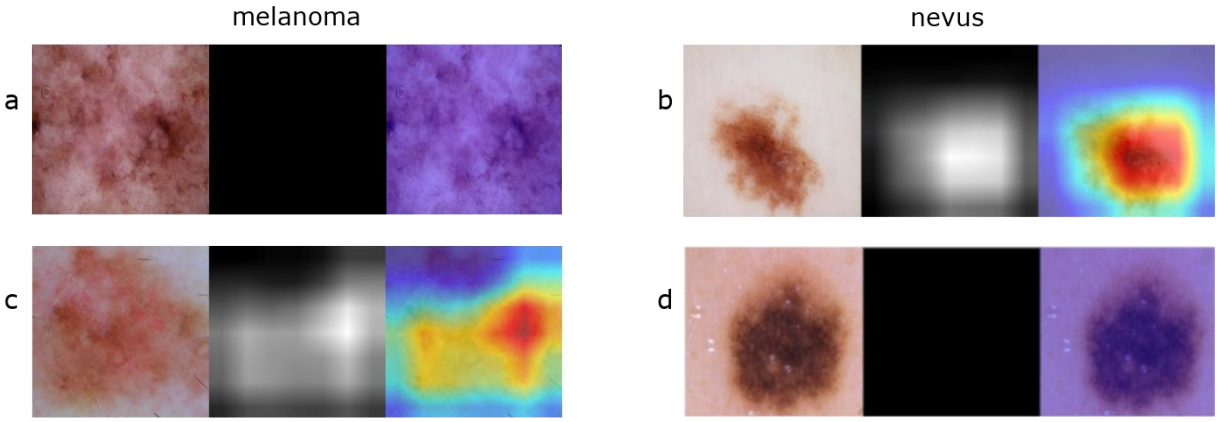

Figure 25: Representative problematic samples from the augmented HAM loc body dataset with resizing, along with their class activation maps (middle image) and their heatmap overlay (right image). The employed model was trained on the augmented HAM loc body training set. a) correctly classified melanoma, b) correctly classified nevus, c) falsely classified actual melanoma, d) falsely classified actual nevus (created with Matplotlib and GIMP 2.10)

Several actual nevus samples classified as melanoma show no activation while correctly classified nevus samples have CAMs that convey plausible activation.

This outcome prompts a further examination of the available saliency methods inside the GitHub[5] repository [35]. Therefore, benchmarks were performed on the non-augmented HAM loc body test set by employing the model trained on the non-augmented HAM loc body training set to find possible deficiencies in the applied saliency method.

Figure 26 illustrates that Grad-CAM's activation is similar to HiResCAM, while Grad-CAM-Elementwise is equal to Layer-CAM. Full-Grad shows identical activation for both classes. This indicates a lack of class discrimination in the implementation of Full-Grad. Overall-Grad-CAM-Elementwise and Layer-CAM produce the most rational activation, since the activation covers the skin lesion. Notably, with every saliency method the prominent nevus bottom bar shape activation pattern is visible.

[5] https://github.com/jacobgil/pytorch-grad-cam

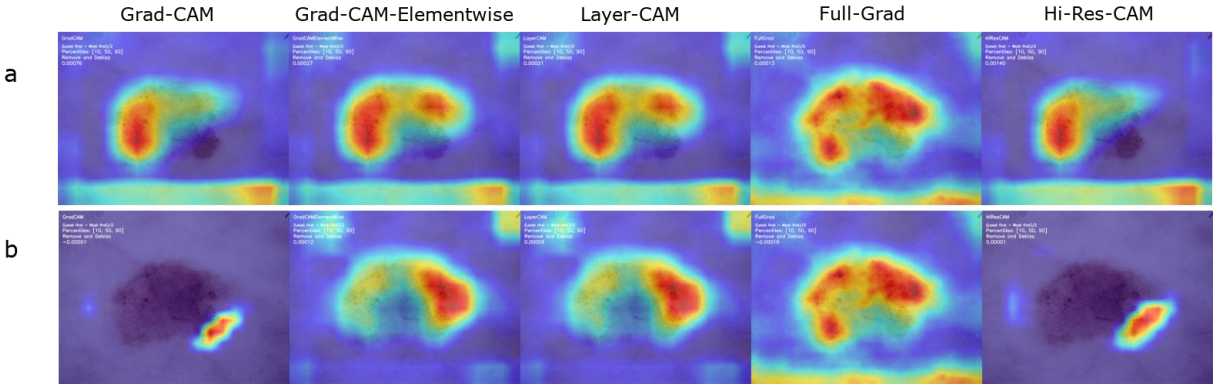

Figure 26: Saliency benchmarks of different methods with a) CAMs for the nevus class and b) CAMs for the melanoma class. The employed model is trained on the non-augmented HAM loc body dataset.

Among the tested methods only Grad-CAM-Elementwise showed both different activation for both classes and focused on the relevant skin lesion area. Therefore, the benchmark is repeated on a sample where Grad-CAM shows almost no melanoma class activation.

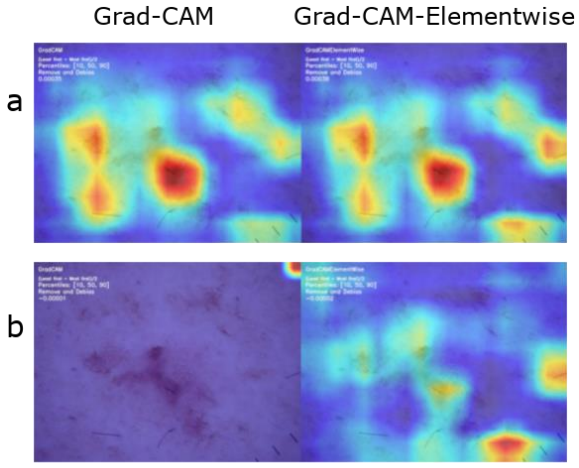

Figure 27: Saliency benchmarks of Grad-CAM and Grad-CAM-Elementwise. a) CAMs for the nevus class and b) CAMs for the melanoma class. (created with Matplotlib and GIMP 2.10)

In the benchmark in Figure 27, nevus and melanoma Grad-CAM activation seem to annul each other in areas with activation for both classes. It is important to emphasize the relative intensity of Grad-CAM at this juncture. Each CAM for a sample is scaled according to the maximum intensity of the activation within that specific sample. Consequently, the color of the heatmap is not an indicator for the absolute activation value in this area. Therefore, areas with low discernible nevus activation could represent a higher absolute value than areas with high melanoma activation. With Grad-CAM-Elementwise this annulling effect cannot be observed.

5.3.3 Augmenting the Source Domain Dataset

Augmenting HAM loc body as the source domain dataset introduces a different testing approach compared to the previous subsections. The effects of augmenting the source domain dataset can only be evaluated by testing the derived model on all other equally augmented domains beforehand. Since this testing approach was not executed so far, this test was conducted first with the model trained on HAM loc body dataset. This model will be referred to as the "original" model. By testing the original model on all domains, a baseline is established for the evaluation of subsequent test results.

Table 8: Test Results of the originally presented model trained on the HAM loc body domain and tested on all other domains. Columns from left to right: employed model, target domain, accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class.

Source Domain	Target Domain	Accuracy	Sensitivity	Specificity	F1
HAM loc body	HAM age under 30	0.866	0.6	0.878	0.286
HAM loc body	HAM loc head neck	0.773	0.65	0.875	0.722
HAM loc body	HAM loc palms soles	0.955	0.33	1	0.5
HAM loc body	BCN loc body	0.71	0.362	0.956	0.508
HAM loc body	BCN age under 30	0.91	0.429	0.951	0.429
HAM loc body	BCN loc head neck	0.551	0.35	0.938	0.506
HAM loc body	BCN loc palms soles	0.517	0.256	1	0.408
HAM loc body	MSK loc body	0.701	0.124	0.965	0.206
HAM loc body	MSK age under 30	0.892	0.143	0.953	0.167
HAM loc body	MSK loc head neck	0.475	0.114	1	0.205

Despite the difference in the testing approaches, one peculiar outcome in Table 8 should be compared to the test results of the domain-specific models on HAM loc body in Table 5. Interestingly, the elevated performance of the models that showed the frame pattern does not manifest when the original model is tested on these domains. In addition, F1-Scores with HAM loc body serving as the source domain are tendentially higher compared to HAM loc body serving as the target domain.

HAM loc head neck stands out with a high performance compared to the other target domains. A prevalently low sensitivity in 8 out of 9 remaining domains is the main reason for the poor F1-Scores of the model. As the performance of the original model is now established, the performance of the model trained on the augmented HAM loc body can be investigated.

Table 9: Test results of the model trained on the augmented HAM loc body domain training set and tested on all other domains. Columns from left to right: source domain, target domain, accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class, absolute change in F1-Score compared to the original HAM loc body model.

Source Domain	Target Domain	Accuracy	Sensitivity	Specificity	F1	F1 Change
aug. HAM loc body	HAM age under 30	0.929	0.6	0.944	0.429	0.143
aug. HAM loc body	HAM loc head neck	0.773	0.6	0.917	0.706	-0.016
aug. HAM loc body	HAM loc palms soles	0.977	1	0.976	0.857	0.357
aug. HAM loc body	BCN loc body	0.749	0.487	0.934	0.616	0.108
aug. HAM loc body	BCN age under 30	0.869	0.5	0.901	0.378	-0.051
aug. HAM loc body	BCN loc head neck	0.631	0.463	0.953	0.623	0.117
aug. HAM loc body	BCN loc palms soles	0.717	0.564	1	0.721	0.313
aug. HAM loc body	MSK loc body	0.776	0.319	0.977	0.465	0.259
aug. HAM loc body	MSK age under 30	0.935	0.143	1	0.25	0.083
aug. HAM loc body	MSK loc head neck	0.475	0.114	1	0.205	0.000

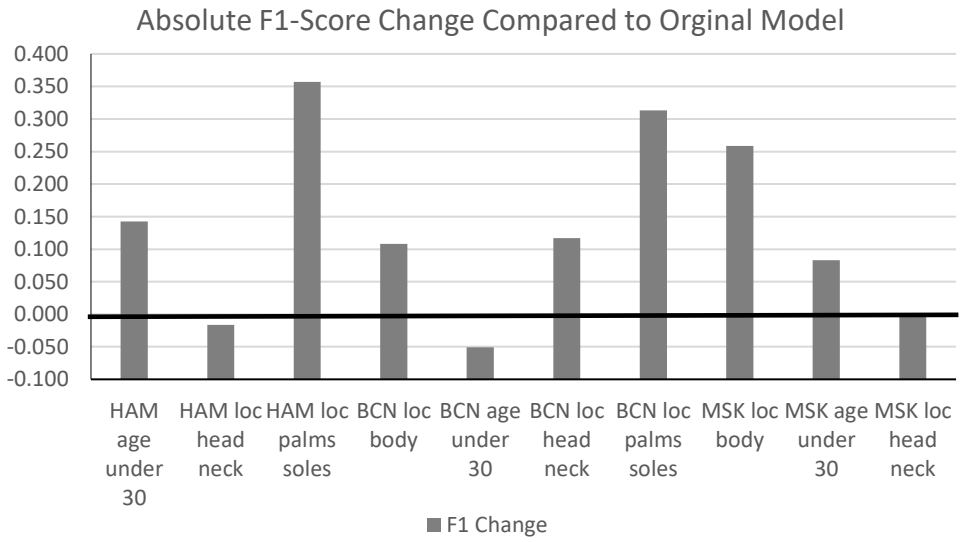

Figure 28: Absolute F1-Score change of the augmented HAM loc body model compared to the original HAM loc body model for every target domain. (created with MS Office)

By augmenting the training set of HAM loc body the derived model was able to improve its performance in 7 out of 10 augmented target domains (see Table 9). The overall F1-Score change is 1.313, signifying a substantial improvement in classification performance. Augmenting the training dataset concurrently improves the overall sensitivity, indicating a more accurate identification of melanoma. Regarding F1-Score, the augmented model now outperforms the original HAM loc body model on the HAM loc body test set in 3 out of 10 target domains. Loc palms soles target domains benefit especially from the augmentation of the training set.

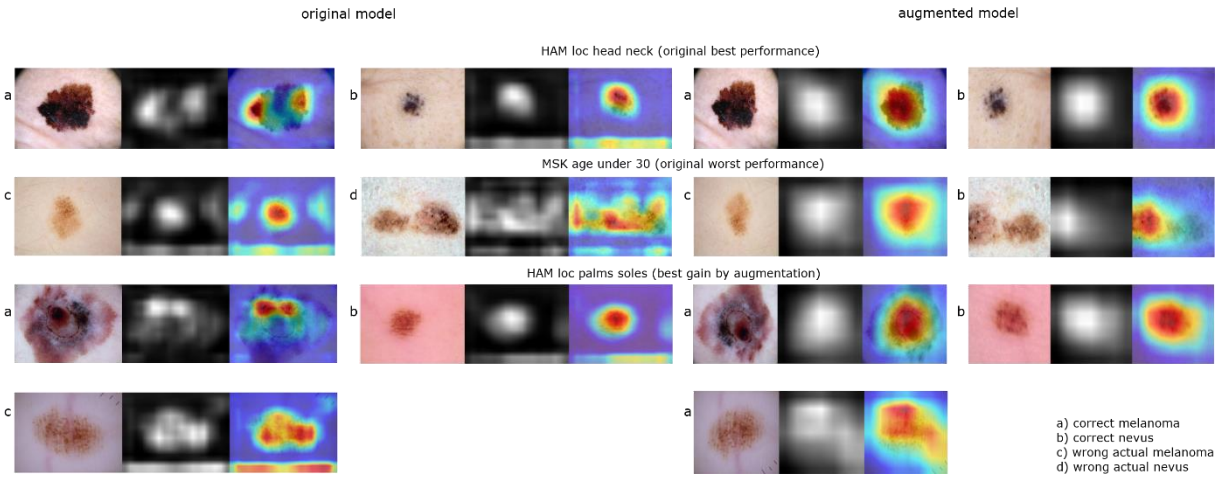

Figure 29: Comparison of representative samples from augmented test sets to their original counterparts of displayed domains, along with their class activation maps (middle image) and their heatmap overlay (right image). Classification outcome for every sample can be derived from the legend. Wrong actual melanoma samples show CAM for nevus class and wrong actual nevus samples show CAM for melanoma class. (created with Matplotlib and GIMP 2.10)

The augmented target domains with the best and worst performance of the original model will be the focus of the examination, as well as the domain with the best improvement. When comparing the CAMs of HAM loc head neck and MSK age under 30 in Figure 29, it becomes evident that the original model applies similar activation patterns to the samples. The model classifies nevus based on a bottom bar activation pattern, while melanoma classification relies on the border regions of the skin lesion. Notably, the skin lesions in the MSK age under 30 dataset tend to be smaller, and unlike the HAM loc head neck model, no red hue is recognizable. Performance on the HAM loc palms soles domain improved significantly due to the correct classification of all melanoma samples. Notably, the augmented HAM loc body model does not show bottom bar activation and the focus of activation is directed at the skin lesions. The activation area tends to exceed the skin lesion area.

5.4 Domain Adaptation on Domain Shifted Datasets

The performance of the domain adapted models on all the target domains is presented in Table 10. It is worth noting that these results were obtained solely by the application of DANN without any augmentation. These results will subsequently be compared to the original and the augmented models' performances.

Table 10: Test results of the DANN models which are tested on the target domain. Columns from left to right: source domain, target domain, accuracy, sensitivity for melanoma class, specificity for melanoma class, F1-Score for melanoma class, absolute change in F1-Score compared to the original HAM loc body model.

Source Domain	Target Domain	Mel. %	Acc.	Sensitivity	Specificity	F1	F1 Change
HAM loc body	HAM age under 30	4.5	0.738	0.6	0.738	0.167	-0.119
HAM loc body	HAM loc head neck	45	0.886	0.8	0.958	0.865	0.143
HAM loc body	HAM loc palms soles	6.9	0.818	1	0.805	0.429	-0.071
HAM loc body	BCN loc body	41.3	0.686	0.534	0.794	0.585	0.077
HAM loc body	BCN age under 30	8.1	0.693	0.857	0.679	0.308	-0.121
HAM loc body	BCN loc head neck	65.7	0.781	0.732	0.875	0.814	0.308
HAM loc body	BCN loc palms soles	64.6	0.833	0.744	1	0.853	0.445
HAM loc body	MSK loc body	44.1	0.692	0.814	0.638	0.617	0.411
HAM loc body	MSK age under 30	8	0.763	0.571	0.779	0.267	0.100
HAM loc body	MSK loc head neck	59.9	0.61	0.686	0.5	0.676	0.471

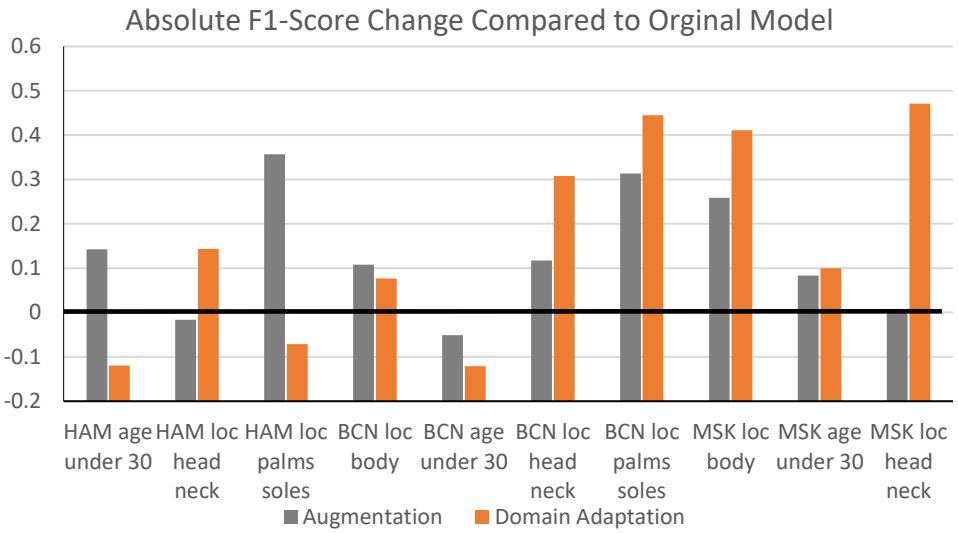

Figure 30: Absolute F1-Score changes of the augmented HAM loc body model and the domain adaptation models in comparison to the original HAM loc body model for every target domain(created with MS Office)

Table 10 and Figure 30 collectively reveal a significant enhancement in F1-Score performance by DANN, demonstrating an overall improvement of 1.643. DANN models now outperform the original HAM loc body model on the HAM loc body test set in 4 out of 10 target domains. However, three domain adapted models exhibit a lower performance compared to the original HAM loc body model. This decrease in performance is particularly noticeable in scenarios where the training dataset contains a low melanoma count. Specifically, a reduction in specificity causes the F1-Score to fall behind the original model's performance.

All domain shifted models demonstrate higher sensitivity on the target domain, whereas this cannot be conveyed to specificity. In fact, specificity decreased in 8 of 10 target domains.

Compared to augmentation, domain adaptation is capable of achieving a higher increase in performance. However, it is important to note that a decrease in Specificity can result in a performance drop below source model effectiveness. This negative impact on specificity, induced by domain adaptation, is notably absent in the results of the augmented models.

Concerning the samples used in Figure 31 one key consideration needs to be addressed. The train/test split function for the target domain was executed without specifying a random state. Consequently, the train and test sets for the target domains were sampled differently in comparison to the original and augmented HAM loc body models. Hence, some previously presented samples now belong to the training set of the target domain. For the purpose of visual comparison, some target domain training samples are presented here, although they did not contribute to the scores on the test datasets in Table 10. For the subsequent quantification however, the same images cannot be provided since the quantification of CAMs is an impartial process, unlike the subjective human interpretation, and would therefore distort the measured values in favor of the domain adapted models, because they have already encountered these samples. When comparing the CAMs of the domain adapted models with the original model in Figure 31, various differences can be identified. For melanoma class activations, the activation covers larger areas of the melanoma but exhibits lower intensity. Nevus activations still emphasize the skin lesion with a tendency to have spots of activation in peripheral areas.

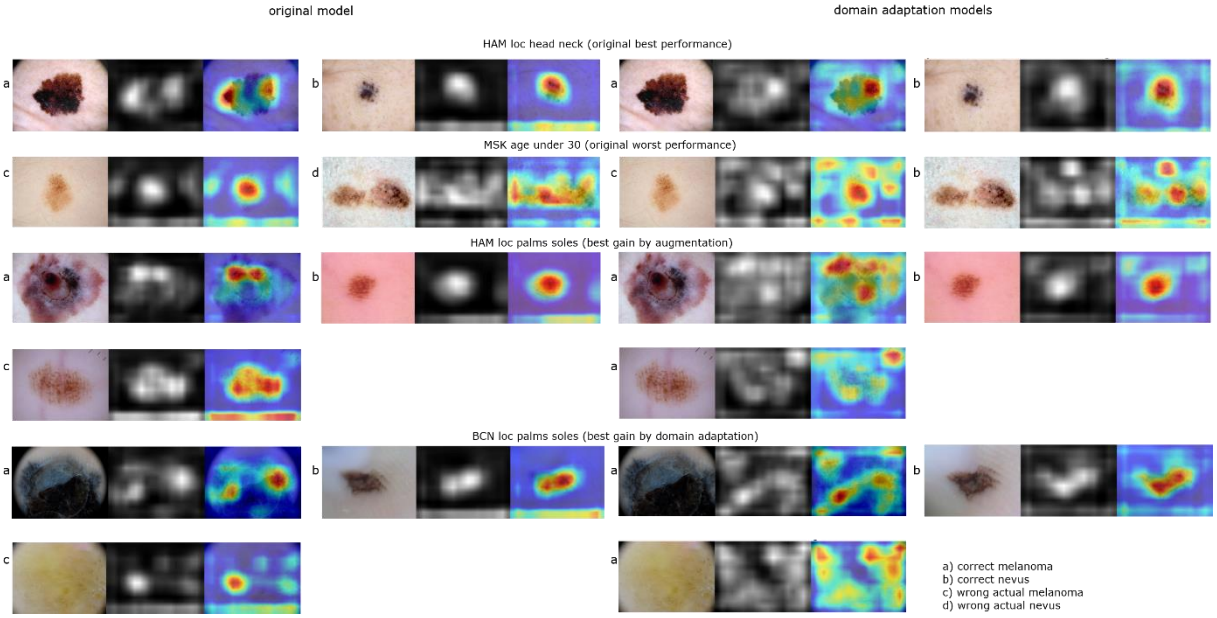

Figure 31: Comparison of representative samples from test sets or training sets of displayed domains, along with their class activation maps (middle image) and their heatmap overlay (right image). Classification outcome for every sample can be derived from the legend. Wrong actual melanoma samples show CAM for nevus class and wrong actual nevus samples show CAM for melanoma class. (created with Matplotlib and GIMP 2.10)

In addition, a frame pattern is noticeable in CAMs for both classes across all domains. This frame pattern has been previously observed when testing the domain models on HAM loc body as the target domain. In comparison to the augmented HAM loc body, the CAMs are less coherent. Nevertheless, the activation remains plausible, since the activation's intensity is high in areas covering the skin lesion. Applying DANN to the BCN loc palms soles domain resulted in a 0.455 increase in F1-Score performance. However, the activation patterns of the model do not show a noticeable difference to the other DANN models.

5.5 Quantitative Comparison of Grad-CAM-Elementwise Class Activation Maps

Prior to applying BCDU-Net to the domain-specific images, it is essential to validate its actual effectiveness on the domain data, since BCDU-Net training was only performed for 10 epochs instead of the 100 epochs proposed in the BCDU-Net publication [27].

Table 11: Performance comparison of proposed BCDU-Net [27] and own BCDU-Net Model, which was trained for 10 epochs.

Method	F1-Score	Sensitivity	Specificity	Accuracy	JS
BCDU-Net (d=3)	0.851	0.785	0.982	0.937	0.937
BCDU-Net (d=3) 10 ep.	0.858	0.825	0.963	0.925	0.751

Evaluating BCDU-Net's performance in Table 11, it becomes apparent that 82.5% of the skin lesion is covered with the predicted segment (sensitivity) and 89.3% of the predicted segment is truly relevant (precision). BCDU-Net achieved an overall Jaccard Score of 0.75 on the ISIC2018 test set, which falls short of the performance specified in the BCDU-Net publication [27]. However, the accuracy aligns with that of the BCDU-Net trained for 100 epochs. Computing the IoU for each image and averaging all IoU values yields 0.745 for the mean. This value closely corresponds to computing the overlap and union from the whole Numpy data frame, suggesting that BCDU-Net performs consistently on large and small skin lesions. An examination of the empirical probability density function of achieved IoU values on the test data frame, separated for images, reveals a distribution skewed to the left. Since 65% of the area under the curve represents 65 % of the BCDU-Net test samples, the threshold for a level of probability of 65% can be derived. The resulting IoU tolerance value is similar to the mean value, and therefore the tolerance can be estimated with 0.26.

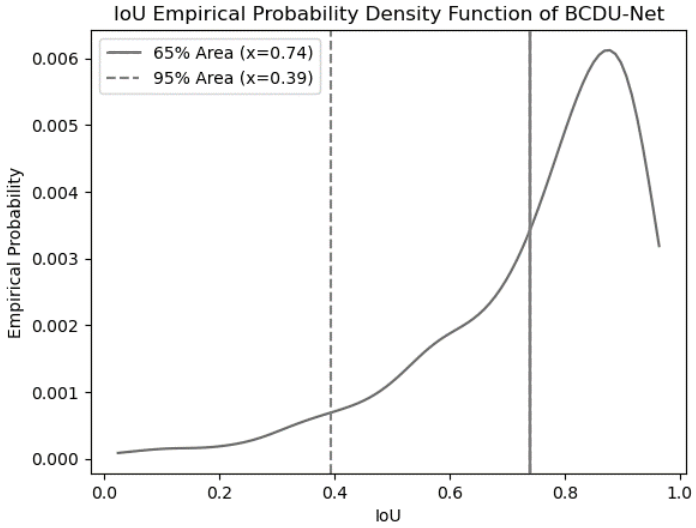

Figure 32:Empirical probability density function of BCDU-Net for IoU values obtained from the BCDU-Net test set. (created with Matlpotlib)

Table 12: F1-Score and mean wIoU in three testing scenarios: Original model, augemented model, DANN models

	Original		Augmentation		DANN	
Target domain	F1	wIoU	F1	wIoU	F1	wIoU
HAM age under 30	0.286	0.226	0.429	0.358	0.167	0.191
HAM loc head neck	0.722	0.195	0.706	0.299	0.865	0.199
HAM loc palms soles	0.5	0.196	0.857	0.27	0.429	0.189
BCN loc body	0.508	0.23	0.616	0.333	0.585	0.238
BCN age under 30	0.429	0.24	0.378	0.329	0.308	0.227
BCN loc head neck	0.506	0.227	0.623	0.304	0.814	0.182
BCN loc palms soles	0.408	0.214	0.721	0.281	0.853	0.214
MSK loc body	0.206	0.186	0.465	0.293	0.617	0.204
MSK age under 30	0.167	0.198	0.25	0.335	0.267	0.193
MSK loc head neck	0.205	0.187	0.205	0.267	0.676	0.186

As Table 12 illustrates, the mean values of wIoU are observed to be the lowest in the DANN testing scenario and second lowest in the original testing scenario. Augmentation led to the most overlap of skin lesions and CAMs with an average wIoU of 0.306 across all domains. In a general sense, an improvement in F1-Score does not necessarily imply an improvement in wIoU. Further, good F1-performance is independent of a high wIoU score as HAM age under 30 and HAM loc head neck emphasize across all testing scenarios.

Table 13: Overview of relative wIoU difference between correctly and wrongly classified samples in augmentation and domain adaptation scenario.

	Augmentation	DANN
Target domain	c/w wIoU relative difference in %	c/w wIoU relative difference in %
HAM age under 30	2.39	-13.86
HAM loc head neck	2.24	-18.85
HAM loc palms soles	-8.99	-4.76
BCN loc body	-1.90	7.23
BCN age under 30	36.57	5.07
BCN loc head neck	-28.15	13.13
BCN loc palms soles	-18.39	1.46
MSK loc body	-3.11	1.10
MSK age under 30	3.93	14.10
MSK loc head neck	-28.12	48.86

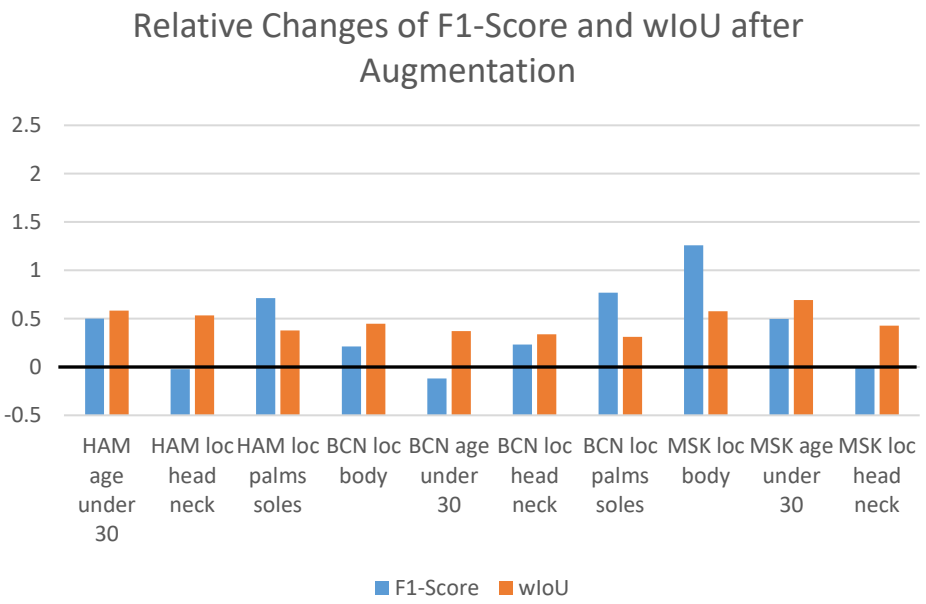

Figure 33: Relative changes of F1-Score and wIoU of the augmented model on all target domains with respect to the original model. (created with MS Office)

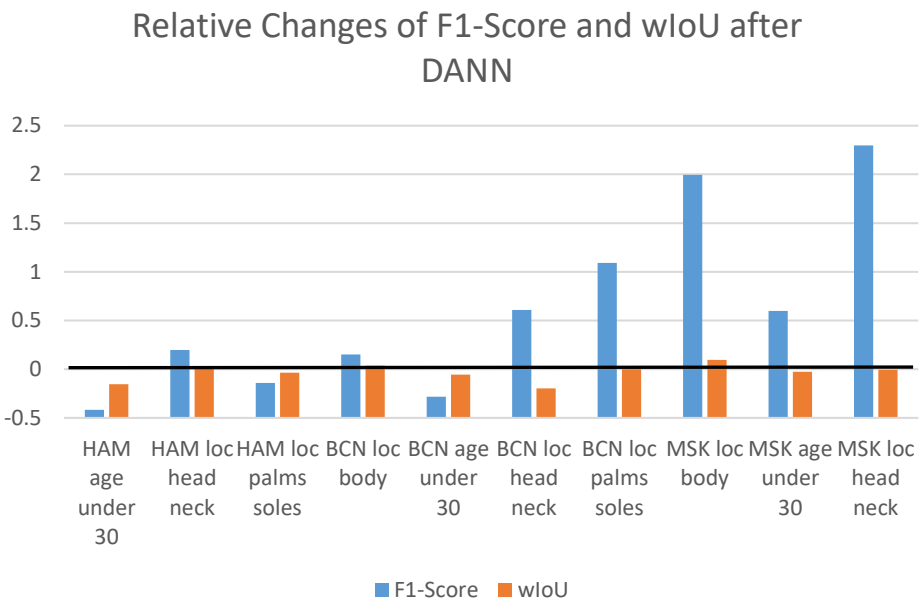

Figure 34: Relative changes of F1-Score and wIoU of the DANN models on all target domains with respect to the original model. (created with MS Office)

Conclusively, no direct association can be identified between the absolute measured values for wIoU and F-Score for both augmentation and domain adaptation. This observation extends to the relative changes in F1-Score and wIoU, as illustrated in Figure 33 and Figure 34. Notably, the relative change of average wIoU of the augmented model never falls below 31 % across all target domains. For the loc palms soles domains and MSK loc body the augmented model exhibits a relative change in wIoU that exceeds the relative change of the F-Score. The original model previously exhibited especially low sensitivity in all of these target domains. It is worth noting that augmentation led to an improvement in wIoU while the classification performance stagnated or even decreased.

Upon assessing the relative wIoU difference between correctly and falsely classified melanoma samples of the augmented model in Table 13, it appears that the correctly classified samples do not exhibit more plausible saliency than the falsely classified samples. However, some domains show a significant difference between correct and wrong samples, surpassing the IoU tolerance threshold. Consequently, the reasons for these deviations are further investigated in Figure 36. Notably, the DANN models exhibited a positive relative difference between correct and wrong classifications across all domains that did not belong to the HAM domain. Most DANN models with a positive relative difference between correct and wrong classifications also exhibited elevated performance on the target domain (see Figure 35).

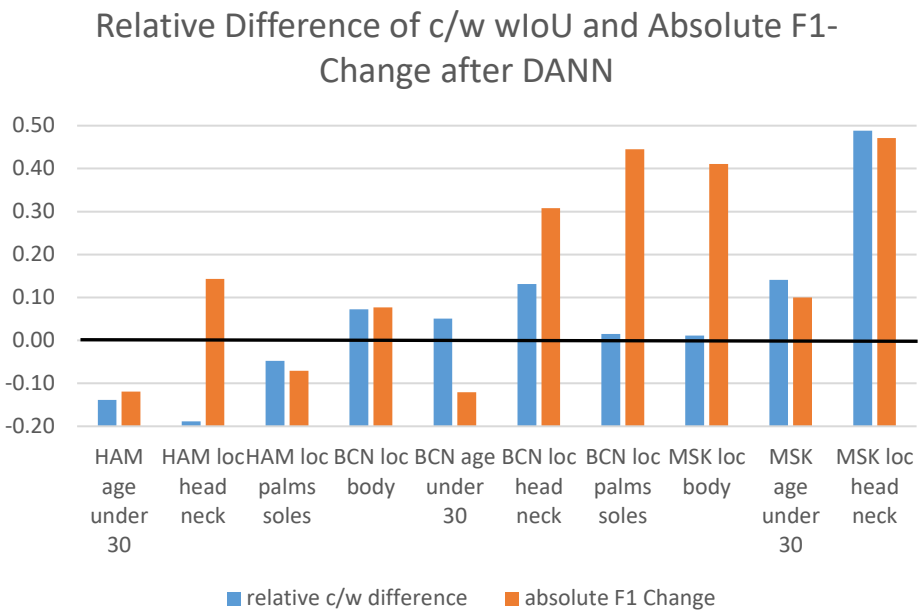

Figure 35: Relative wIoU difference between correct and wrong classifications and absolute F1-Change of DANN models on all target domains (created with MS Office)

Examining samples from correctly and incorrectly classified BCN age under 30 images reveals that smaller lesions tend to be overrepresented, accompanied by a widespread activation area (see Figure 36). This leads to a lower wIoU score for the falsely classified samples. For the DANN testing of the MSK loc head neck, similar principles apply, since the lesions overlap more with the frame-like activation pattern of the model. This pattern is reversed in the case of BCN loc head neck and MSK loc head neck.

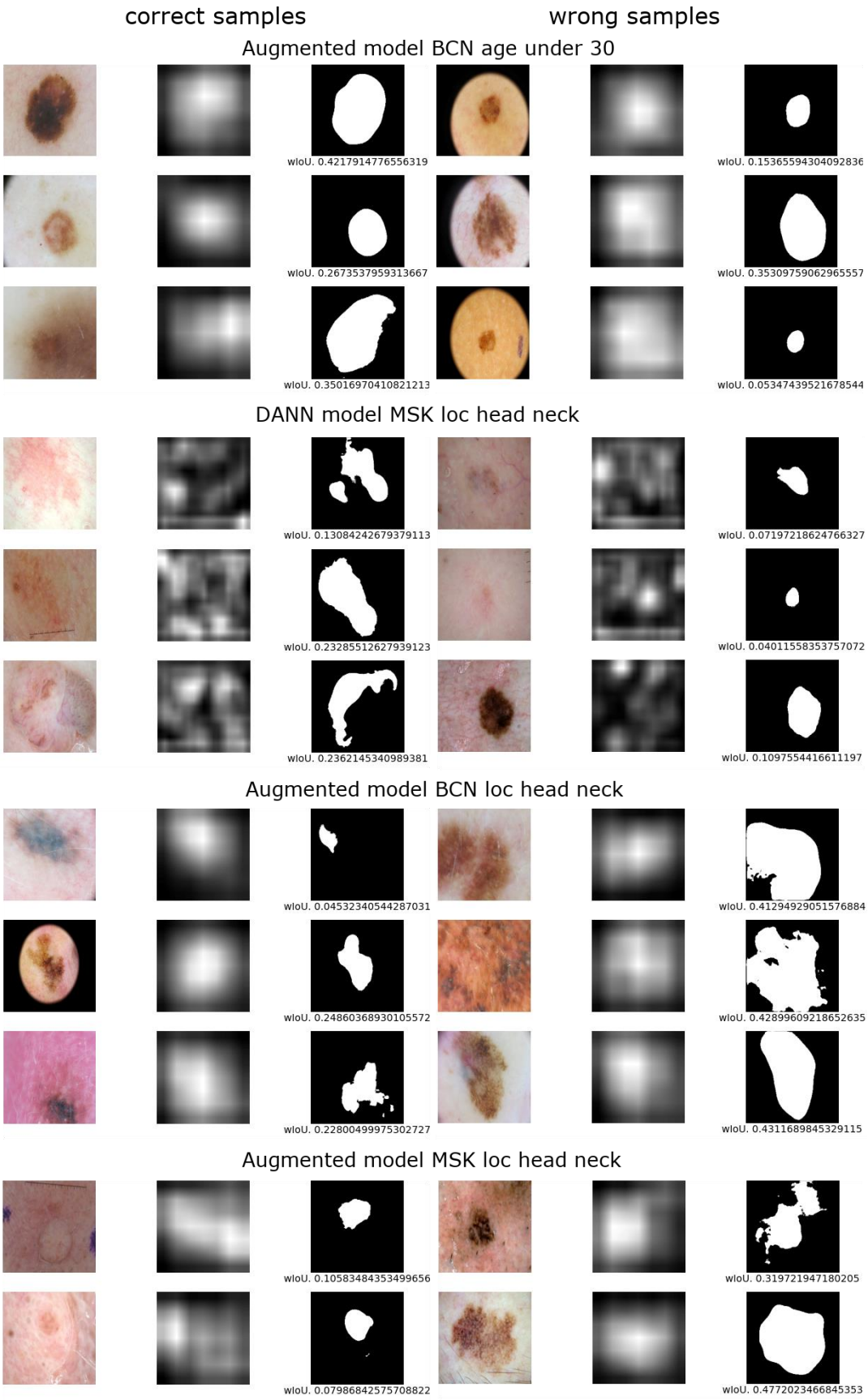

Figure 36: Correctly and falsely classified samples of various domains (left image) along with their class activation map (middle image) and thier skin lesion segment (right image) (created with Matplotlib and GIMP 2.10)

6. Discussion

6.1 Saliency on a Multi-Domain Dataset (Inter-Domain-ISIC Dataset)

Before domain shifts can be explored with a ResNet18 model, it is essential to establish a benchmark and compare it with state-of-the-art approaches. This benchmark provides a baseline against which deviations from typical values and activation patterns can be identified. Rehman et al. [58] previously explored binary skin lesion classification and provided Grad-CAM outputs with a comparable dataset. Despite the smaller and less balanced training dataset, Rehman et al. have achieved an F1-Score of 0.94 on ISIC images using a DenseNet201. Notably, Rehman et al. employed extensive image augmentation and more complex neural networks. Despite the renouncement of the extensive augmentation approach, it is apparent that this ResNet18 model does not reach state of the art performance. Nevertheless, this performance result serves as a valuable reference point to evaluate effects of domain shift on performance of ResNet18 models.

Regarding the activation patterns, Rehman et al. also explored Grad-CAM activation of DenseNet201 on five samples. Rehman et al. did however not focus on differentiation between the class specific activations or correct classification outcome. They present Grad-CAM output that illustrates the model's capability to focus on the skin lesion, without specifying the type of skin lesion or the specific activation patterns visible in the sample. Their result can be confirmed here since the focus of the activation of ResNet18 is also on the skin lesion except for cases of wrongly classified actual nevus samples. From the identified CAM patterns can be derived that in models, trained on an inter-domain dataset, Grad-CAM output can prove advantageous for determining the correctness of nevus classification.

6.2 Interpretation of Statistical- and Activation-Differences between Domains

Since domain shifts inside the ISIC dataset were quantified [10] it is of interest how CNN activations change and what insights can be derived from them as domain shift occurs. Specifically, the failure modes and beneficial particularities of domain-specific models are relevant.

Before evaluating the performance of domain-specific models on the HAM loc body domain, it is essential to first assess the performance of a ResNet18 that was trained on this domain. Surprisingly, the F1-Score of the classifier is observed to be lower than that of the previously discussed inter-domain model. Since the data variability inside a domain is expected to be

smaller than that of an inter domain training set, this outcome may appear counterintuitive. The training process of the HAM loc body, also previously referred to as the original model, only differs from the inter-domain model in its preprocessing. Resizing however reduces the available amount of information. Renouncing resizing should therefore be beneficial to the performance of the original model.

In contrast to the original model, all domain-specific models exhibit great performance on their according test set, with the exception of MSK loc head neck, which still outperforms the original model. During training minor instances of overfitting can have occurred, since some domains contain a low number of samples [10]. In this context, the normalization of the training data helps mitigating the adverse effects of overfitting [63]. These results are expected and show that training the domain-specific models has been successful.

The examination of the F1-Score performance of the domain-specific models has revealed an allocation between 5 better performing models and 5 poorer performing models. From the statistical data no distinct impact factor can be identified that can be accounted for the failure of some domain-specific models. Therefore, analyzing the image data and CAM patterns of the domains is necessary to gain further insights in the underlying factors contributing to both superior and inferior performance.

CAM patterns of the original model indicate that a bias is present in the HAM loc body training dataset, since the model learned to classify nevus images based on the bottom portion of the image. This bias could impede a reliable recognition of melanoma samples within the test set. When the bottom portion of a melanoma sample shares similarities with a nevus image, the model may incorrectly classify it as a nevus. A Comparison of all available samples further indicated a bias in the hue of the images, since the model was more successful in classifying images with a red hue. Therefore, the training dataset of the HAM loc body domain has been examined for a homogenous bottom area and a red hue in the nevus samples. The investigation of the random sample affirmed the presumed biases. This presents a suitable explanation for the subpar performance of the original model on its own test set.

Since the HAM loc body dataset contains color and location biases, and considering that the activation maps are influenced by these biases, it becomes necessary to investigate whether the improved performance of certain domain-specific models can be attributed to similar activation patterns. The HAM loc head neck model demonstrates significantly different activation patterns depending on the classification outcome. These patterns remain consistent, as the model is tested on the target domain. The activation patterns and the performance bear a resemblance to

the original model. This indicates a link between the similarity of CAM patterns and the resulting performance on the target domain.

The examination of the CAMs of the BCN loc head neck model revealed similar patterns and observations to those found in HAM loc head neck. This adds to the presumption that similar patterns are manifesting in higher performance. So far, the best performing models have been trained on a loc head neck body region. This could lead to the assumption that the domain shift between these domains and HAM loc body is smaller, potentially resulting in improved model performance. The cosine similarity between nevus images of HAM loc body and the target domains states that HAM loc head neck contains the most dissimilar nevus samples inside the HAM domain [10]. Furthermore, BCN loc head neck nevus images are least similar of all the domains. The trend could however be reversed, as models trained on domains with low nevus similarity to HAM loc body tend to exhibit higher sensitivity values. This phenomenon might be attributed to the binary classification task, in which the derived models would tend to decide for melanoma, resulting in elevated sensitivity and a positive influence on the F1-Score.

Moreover, the MSK loc body model as the third best performing model challenges the assumption that HAM loc body shares similarities with the loc head neck domains. In all previously examined domains, CAM patterns of domain specific models remained relatively consistent when confronted with target domain images. The MSK loc body model however unexpectedly shows a tendency for bottom-bar or frame-shaped nevus CAM patterns on the target domain. Unfortunately, no indicators for this phenomenon can be identified. This undeniably proves the thesis that a neural networks activation pattern varies as the domain is changed within the ISIC-Dataset. So far this is observation is exclusive for the MSK loc body model and indicates an elevated generalization capability.

For all good performing domain specific models, a consistent frame or a bottom bar pattern can be identified for nevus CAMs in the source domain. Interestingly, this pattern either gains intensity or establishes when the model transition to the target domain. This observation indicates that the presence of a peripheral activation pattern is advantageous in testing on the HAM loc body dataset.

Among the five poorly performing domain-specific models, BCN loc body stands out as a dataset with a balanced melanoma share and a large number of training images. Based on statistical values, models trained on BCN loc body would be expected to exhibit good performance on the HAM loc body test set. Consequently, the CAMs of this model are particularly significant for the identification of failure modes. When confronted with HAM loc body test samples, the BCN loc body model’s CAMs suffer a decrease in plausibility. Notably,

the previously observed frame pattern, which was prominent in better-performing models, is conspicuously absent in cases of low performance. The observed overrepresentation of samples with red hue in the misclassified samples is an additional indication for the low performance on the target domain.

These observations indicate a connection between performance on the HAM loc body test set and the presence of a frame or bottom-bar nevus CAM pattern. However, it is worth noting that no indicative patterns can be derived from the melanoma CAMs since they are spot-wise and sample dependent. These findings prompt the development of a theory that postulates the existence of underlying concurrent location bias in the training datasets of some domains. To substantiate this theory, the bias needs to be eliminated from the target dataset before the performance and CAMs of the successful domain-specific models is reassessed. If the removal of the bias leads to a significant performance drop in the five better performing models, then the theory can be considered as substantiated.

6.3 Effects of Augmentation on Performance and Activations

6.3.1 HAM Loc Body as the Target Domain

As elucidated before, the augmentation of the target domain is an essential step in proving the concurrent biases theory. Therefore, the examination of performance with respect to the corresponding CAM patterns is deemed to provide insights.

The random augmentation of the nevus images did not decrease the performance of domain-specific models that were trained on domains where concurrent bias is presumed. Surprisingly, the frame pattern, characteristic for nevus activations, is still discernible across all domains that exhibited better performance on the non-augmented HAM loc body test set. The illustrated association between performance on the non-augmented HAM loc body dataset and the intensity of the nevus frame-pattern of the CAMs in Figure 23 needs to be further examined. As the cropping of the images did not successfully eliminate the size difference between the melanoma and nevus samples, the augmentation was unable to address this aspect of the presumed biases. Since no tendency for an intense bottom-bar activation pattern is discernible in the nevus samples, this bias seems to have been successfully eliminated.

Based on these findings, it appears that performance is not strongly associated with the homogenous bottom region of the nevus samples. Instead, it is plausible that performance is linked to the peripheral regions in the nevus samples Considering that the domain-specific model was trained on images with a particular aspect ratio, it becomes evident that this specific

factor needs to be addressed to dissipate the aspect ratio-driven frame pattern. This prompts the additional resizing within the augmentation process of the HAM loc body test images.

In the experiment with the resized images, the frame patterns on nevus samples were successfully eliminated. The corresponding statistical test results support the theory that the frame pattern is responsible for the elevated performance on the non-augmented HAM loc body test set. Despite the significant difference between the resized input images and the initial full-scale training images, only domains with a presumed concurrent bias experienced a significant drop in performance. Unexpectedly, the HAM loc palms soles dataset gained 0.271 in F1-Score performance. Although the HAM loc palms soles model exhibited a frame-shaped activation pattern for nevus, the activation area still covered larger areas of the skin lesion. Speculatively, this phenomenon may have contributed to the improved performance of this specific model, even though the formation of the frame pattern was averted by augmentation.

6.3.2 The Most Suitable Saliency Method for Quantification

An unexpected deviation from the previous results emerged when a ResNet18 model was trained on the augmented HAM loc body dataset. As depicted in Figure 25, the implausible Grad-CAM activation cannot be attributed to model failure or other factors. It is unlikely for a well-performing model to exhibit zero activation for one class, while deciding against the class with rational activation. In pursuit of quantifying saliency, a reliable saliency method has to be ascertained.

Among the examined saliency methods solely Grad-CAM-Elementwise has exhibited sane activations for both classes. Therefore Grad-CAM-Elementwise was directly compared to the previously employed Grad-CAM. The benchmark results indicate that Grad-CAM is unsuitable for quantifying the impact of augmentation and domain adaptation on CAMs within domain shifted datasets.

The observation of potential annulation of activation in Grad-CAM is presumably attributed to the close proximity of the features in the fourth layer of the ResNet18 model, the layer to which Grad-CAM is applied. This proximity issue is compounded by the training of the model on a strongly imbalanced two-class dataset, where feature maps are likely to predominantly represent features of the overrepresented class. Since melanoma features and nevus features should be close nearby, they are inevitably superimposed within the standard 7x7 resolution of the layer 4 feature maps. Consequently, when Grad-CAM or HiResCAM aggregates all feature maps, this consolidation could result in the vanishing impact of melanoma feature maps.

Grad-CAM-Elementwise applies ReLU before summing the elementwise product of gradients and feature maps. This approach disregards the influence of negative product values, allowing the contribution of melanoma feature maps to persist. As the only difference between Grad-CAM-Elementwise and HiResCAM is the application of ReLU activation before summation, this explanation appears logically sound.

The issue did not occur before, because melanoma and nevus activation were not as precisely superimposed or class imbalance was less severe. The alignment of activations occurred here due to the augmentation of training data. The augmentation led to a focus on the skin lesion, as it randomizes other features and effectively mitigates biases.

6.3.3 HAM Loc Body as the Source Domain

The augmentation of HAM loc body as the target domain initially focused on verifying biases and did not benefit the classification performance of the domain-specific models. However, with the introduction of data augmentation within the HAM loc body domain, the focus has transitioned towards seeking performance improvements. In this testing scenario, the effect of augmentation on the CAMs is particularly interesting in view of dataset bias mitigation.

A baseline is necessary to evaluate effects of augmentation or domain adaptation in regards of performance as well as qualitative and quantitative CAM analysis. For this reason, testing of the original model was conducted on all target domains.

The analysis of Table 8 yielded the that the previously observed elevated performance of some domain-specific models on the HAM loc body test set is not reciprocal This discrepancy can likely be attributed to the homogenous bottom area that the original model focuses on, when predicting nevus. This bottom area is not exclusive to nevus samples within the target domains. Hence, the original model exhibits tendentially low sensitivity on the target domains.

Performance on the HAM loc head neck test set is the highest, which coincides with the HAM loc head neck model performing best on the HAM loc body test set. One potential explanation lies in the dissimilarity of nevus images [10], which could result in higher sensitivity, especially if melanoma images exhibit similarities. Conversely, the lowest performance was observed on the MSK age under 30 domain's test set. Upon analysis of the dataset and the corresponding activations it became evident that the skin lesions in this particular dataset are tendentially smaller and some incorrectly classified actual melanoma contain a red hue. These factors could be the primary contributors to the low performance of the original model.

The established baseline performance, enables the comparative evaluation of the augmentation's effects on performance. From Figure 29 it becomes apparent that the bottom

bar pattern of the original model has been eliminated by source domain augmentation. This pattern certainly originated from the previously discussed dataset bias inside the HAM loc body training dataset, where uniform bottom regions were clearly overrepresented in nevus samples. This bias induced nevus activation pattern led to low sensitivity of the original model across all target domains, consequently negatively influencing the F1-Score performance. The successful elimination of the nevus bias and the associated bottom bar activation pattern has had a positive influence on sensitivity and performance in 7 out of 10 domains.

Prior to augmentation, actual melanoma samples within the loc palms soles domains were incorrectly classified as a nevus because of the bottom bar activation pattern. Consequently, the removal of the activation pattern has proven particularly advantageous for enhancing performance within the loc palms soles domains.

6.4 Comparison of Statistical and Activation-Differences between Augmentation and Domain Adaptation

6.4.1 Statistical Comparison

Since DANN is one of the most well-established approaches [45], it is especially interesting to compare it to the performance enhancement that augmentation yielded. With respect to the nevus bias inside the HAM loc body dataset being one of the primary reasons for the original model's failure on the target domains, it is particularly interesting to see how domain adaptation without augmentation handles the biased source dataset.

As expected, DANN was able to improve performance more than the extensive augmentation of the training data. Unexpectedly, DANN worsened performance in 3 out of 10 domains. This circumstance is not entirely new, since the low gain in performance or stagnation was observed in association with low melanoma share in the target domain in previous works [10], [64]. Therefore, it is plausible that the tendentially lower sensitivity of the DANN models impacts the F1-Score performance more in domains with overrepresented nevus samples. The results, presented in these works, were obtained by averaging over five seeds. This could explain the distinct decrease observed on HAM age under 30 and BCN age under 30. However, the domains that benefited most from the domain adaptation in other works do not coincide with the domains with increased scores. For this, no underlying cause can be identified. It is however possible that the different backbone network led to different performance [64].

In direct comparison to the augmentation approach, the augmentation is preserving the specificity of the classifier. This does not significantly lessen the classifiers performance on domains with low melanoma share. Notably, both approaches managed to outperform the

original HAM on different target domains. Upon comparison of domains which now show excellent performance, it appears that augmentation and domain adaptation complement each other in this regard. This sparks the idea that extensive augmentation and domain adaptation could compensate their shortcomings. Further details of this approach are discussed in the Outlook section.

6.4.2 Grad-CAM-Elementwise Activations

Previous studies have demonstrated the effectiveness of unsupervised domain adaptation (UDA) techniques in dermoscopic datasets within the ISIC domains [10], [64]. Here, the impact of DANN on network activations is explored. The objective is to examine whether new insights into the influential factors of UDA can be obtained through the analysis of CAMs generated by DANN models.

Figure 31 reveals a significant alteration in activation patterns across the domains due to domain adaptation. The previously observed nevus frame shape pattern, distinctive to the domain-specific models, is discernible with varying intensity in the CAMs for both classes across all domains, as Figure 31 illustrates. In some domains, the lower portion of the frame pattern exhibits higher intensity, indicating that the bottom-bar bias has not been thoroughly removed. However, the overall perfromance improvement achieved with domain adaptation surpasses that of augmentation. This indicates that this bias is not necessarily disadvantageous for classification performance of DANN models.

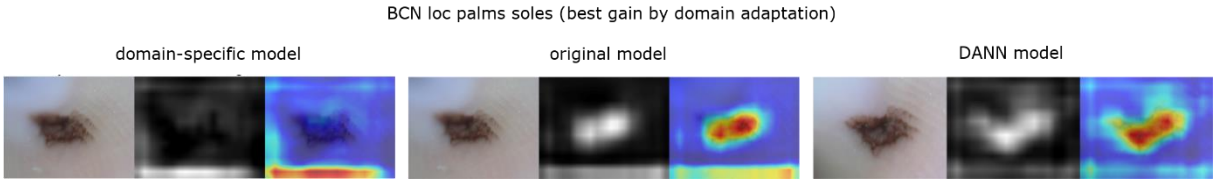

Figure 37: Example of the reintroduction of the frame shape activation to CAM patterns of the DANN models. The presented sample is a nevus (left image), classified correctly by all displayed models along with the CAM (middle image) and its heatmap overlay (right image). (created with Matplotlib and GIMP 2.10)

Interestingly, the patterns established by DANN remain mostly consistent regardless of the domain. This suggests that, across all domains, the background plays a role in accurately classifying samples inside the target domains. Remarkably, the reintroduction of the domain specific activation patterns occurred without access to the target domain labels (see Figure 37). Since DANN's optimization objective is to align source and target domain feature vectors to be indistinguishable, this pattern emerged without the utilization of labels. This marks the initial point of a potential explanation, concerning dataset imbalance as a primary influential factor in unsupervised domain adaptation.

Within the process of unsupervised domain adaptation, the influence of image features from the target domain on the resulting DANN model depends on their relative frequency. In case of an underrepresentation of melanoma samples, the majority of the influence originates from nevus samples. Therefore, the resulting feature extractor provides the classifier with features that predominantly resemble those of nevus samples. In the process, actual melanoma samples of the source domain are modified so that their resulting feature vector is more similar to a nevus feature vector. As a result, the classifier learns to recognize melanoma based on patterns previously associated with nevus. According to the observed CAM patterns, the classifier noticeably decides for melanoma based on the nevus frame pattern observed in the source-domain models. This leads to higher sensitivity on a balanced target domain test set. Conversely, in an unbalanced target domain test set, the tendency for the melanoma class is unfavorable for the specificity, increasing the likelihood of misclassifying an actual nevus as melanoma. It is important to emphasize that this explanation is solely based on the observation of the CAMs and requires further research and experimentation for validation. This explanation is primarily applicable to a binary classification problem and limited to the DANN architecture. Regardless of the particular explanation regarding the source of class imbalance performance issues, a significant conclusion can be drawn. The previously discussed results, emphasize the importance of conducting a previous assessment for class imbalance by experts prior to unsupervised domain adaptation. In case of identified severe class imbalance, other proposed domain adaptation approaches like a Feature and Label Distribution Co-Alignment Model (COAL) [65], which are explicitly designed to address class imbalance, have to be considered. Concerning other influential factors on UDA, mentioned by Chamarthi et al. [64], no additional insights can be gained by the examination of CAMs. Domains with high melanoma share also consistently exhibited improvement, while no substantial difference in activation patterns could be identified. Age under 30 domains were also found to pose challenges for adaptation, yet no clear indicators could be determined from the activation difference. Large datasets have been found to be advantageous, which the statistical results have affirmed. The according CAMs revealed a focus on skin lesions, though this characteristic is not exclusive to large datasets. Consequently, the insights obtained by assessing Grad-CAM-Elementwise activations remain very limited in this context.

6.5 Evaluating Effects on Saliency of Augmentation and Domain Adaptation

As dermatologists use various diagnostic criteria, it presents a challenge for non-specialists to determine the most plausible CAM for a specific sample. Nonetheless, dermatologists predominantly focus on variables inside the skin lesion to recognize melanoma [66]. Therefore, activation can generally be regarded as plausible in case the CAM activation overlaps with the skin lesion. Recent research has explored the plausibility of Grad-CAM activations in dermoscopic images, accompanied by subsequent attempts to quantify their plausibility.

Gamage et al. [67] employed Grad-CAM and ISIC2018 cell network segments, referred to as attribute masks, which only encompass portions of a complete skin lesion segment. They aimed at gauging neural network activation for their plausibility. Their investigation revealted that Grad-CAM++ exhibited better coverage of the attribute masks than Grad-CAM, although the same network has been employed. Furthermore, they suggested the use of IoU for quantification as a possible extension of their work. Their findings align with the previously discussed results, underscoring the significant impact of the chosen saliency method on CAM results and their plausibility.

Further studies by Lee et al. [68] and Nunnari et al. [69] quantified the alignment of skin lesion segments and Grad-CAM activations in the context of differing network architectures. They tried to determine which network architecture exhibited a stronger focus on critical regions inside the images. In both studies the authors decided for a binarization of the CAMs along with an optimization of the threshold. Lee et al. proved that their proposed network architecture was able to achieve 0.05 higher IoU. Notably the classification performance difference was only 0.011. These findings therefore do not necessarily contradict the previously presented results of unassociated wIoU and F1-Score. Nunnari et al. illustrated that VGG16 is showing more plausible activation in correctly classified melanoma than in incorrectly classified actual melanoma. In contrast, ResNet50 CAMs were not able to provide this benefit, despite better classification performance. They concluded that the higher resolution saliency maps provide greater explanatory value for dermatologists.

As the available research on quantification of Grad-CAM activations is limited, wIoU was established as a custom approach in order to gain information about the overall plausibility of activations. The quantification of CAMs using a wIoU metric is unique in research until this point.

Because no model weights for the skin lesion segmentation task were provided by the BCDU-Net GitHub[6] repository [62], the model had to be retrained. However, due to the tedious training process, the training has been finalized after 10 epochs. According to the BCDU-Net publication [27], comparable validation performance was already achieved after 10 epochs of training. Since the training did not extend to 100 epochs, an evaluation of the BCDU-Net on its test set was necessary to assess the performance of the BCDU-Net. A detailed examination of the results led to the conclusion that the tolerance range of the IoU is estimated to be 26%. Therefore, all relative changes in wIoU beyond 26% are considered significant.

Table 12 reveals that no meaningful connection between performance and wIoU can be discerned for either augmentation or domain adaptation. This unexpected finding suggests that a high-performing model does not necessarily prioritize the skin lesion. Notably, augmenting HAM loc body as the source domain resulted in the highest overall wIoU. This underscores the independence of wIoU and classification performance, since augmentation did not yield higher performance improvement than DANN. As discussed previously, the presence of bias within the training datasets for DANN models does not appear to be disadvantageous for good performance.

The bias-induced activation pattern of the original model, which predominantly led to the misclassification of melanoma, likely explains why augmentation led to improved wIoU across all target domains, even as classification performance remained stagnant or declined. Augmentation effectively removed this bias, causing the model to consistently prioritize the skin lesion This resulted in significant performance enhancements, particularly in domains with previously low sensitivity.

It is necessary to emphasize the crucial role of explainability in the application of AI in clinical scenarios. Therefore, it is of great interest whether wIoU of correctly and incorrectly classified images are different to each other. Specifically, a dermatologist should be able to discern correct and wrong classification based on the sensibleness of a CAM.

When analyzing the Results in Table 13, it appeared that correctly classified samples do not exhibit more plausible saliency than the incorrectly classified samples for augmented models. This observation coincides with the findings of Nunnari et al. [69] and challenges the idea that dermatologists can trust neural networks in case of high plausibility of CAMs in diagnosing melanoma.

[6] https://github.com/rezazad68/BCDU-Net/tree/master

For DANN models, all target domains outside HAM exhibit better wIoU on correct than on wrong samples While most of these differences in wIoU metrics fall within an acceptable tolerance range, this pattern warrants further investigation due to the clear separation observed between HAM and other domains.

This particular segregation is accompanied by a more substantial improvement in classification performance for target domains beyond HAM. However, it's important to note that this observation may be attributed to the influence of class imbalance. Within this context, it becomes evident that HAM domain-specific models lack a focus on skin lesions. Specifically, they tended to exhibit no activation on critical regions in nevus and minimal activation on melanoma samples. The domain adapted models therefore expressed a lack of plausible activation and focused on peripheral regions to classify the images. This behavior appears logical since these peripheral features were suitable to high performance on the target domain. It's worth emphasizing that this phenomenon is distinct from the previously discussed frame pattern, as even the BCN loc head neck DANN model demonstrates a higher wIoU for correctly classified samples.

Figure 37 illustrates this phenomenon, where the BCN loc palms soles domain-specific model exhibits low activation on skin lesions, resulting in a limited wIoU advantage for correctly classified samples, despite a great improvement in F1-Score, for the corresponding DANN model. This observation indicates that the CAMs of domain adapted models may only provide practical benefits for the diagnosing dermatologists in case the target domain does not contain bias, influencing the DANN model to focus on the skin lesion. However, it is important to acknowledge that it is challenging to identify whether the target domain contains bias that would influence the model's focus on skin lesions prior to domain adaptation.

Additionally, some target domains exhibit significant difference in wIoU between correctly and wrongly classified samples, which are associated with discrepancies in skin lesion sizes. The observed significant differences are therefore not relevant for the idea of assistive saliency maps in melanoma diagnosis.

These findings underscore the findings of Nunnari et al. [69] and contest the ability of saliency maps of ResNet models to effectively assist dermatologists in diagnosing melanoma or help establishing trust in ai based skin cancer classification. Across all examined scenarios, saliency maps of neural networks do not exhitbit significant difference in wIoU between correctly and incorrectly classified melanoma samples. The sole exception is observed in the subjective CAM analysis of the inter-domain model.

In conclusion, the combination of Grad-CAM-Elementwise with domain adapted or augmented ResNet models is not suited for establishing trust in AI-based skin cancer classification. Nevertheless, the application of saliency maps in domain adapted or augmented models is best suited for model examination, enabling the recognition of patterns in activations and the identification of biases in the training data. Therefore, it also serves as a valuable identification tool for failure modes in cases of low model performance.

7. Conclusion

The results and their interpretation enabled further insights into influential factors of domain shift and the behavior of saliency methods in domain shift scenarios. In accordance to the initially formulated objectives, the primary findings of this work shall be summarized.

The initial findings and their interpretation have revealed the potential advantage of inter-domain models' Grad-CAM activations for dermatologists in diagnosing skin cancer.

The examination of causes for domain shift yielded that domain-specific models tend to perform well on each other's test sets when they exhibit similar activation. This observation emphasizes that Grad-CAM allows insights regarding failure modes or advantageous patterns within the domain training data or the trained models. Specifically, location and hue biases were identified in the HAM loc body domain, along with concurrent location and or hue biases inside other domains. Additionally, it became apparent that biases inside the source domain exert a dominant influence on the CAM patterns observed in the derived model.

It is important to note that significant changes in Grad-CAM activations were exclusively observed with the MSK loc body domain-specific model as the target domain was changed. This observation confirms that domain-specific models generally have a limited generalization capability.

During the experiments it became evident that Grad-CAM output is not suited for quantification of CAMs in binary and domain specific skin cancer classification tasks. However, Grad-CAM-Elementwise guarantees plausible activation maps. These findings advise caution when employing Grad-CAM in binary skin cancer classification tasks. The use of Grad-CAM-Elementwise, however, was found to be more reliable.

The extensive augmentation applied to both the source and target domain changes the activations significantly and concurrently improves classification performance. Therefore, extensive augmentation is an effective tool in mitigating geometric bias and is strongly

recommended in order to obtain models with enhanced generalization capabilities, capable of adapting to previously unseen domains.

Domain adaptation uses dataset bias to its advantage to optimize performance on the target domains. Subsequently, CAMs do not necessarily show plausible activations in models with superior performance on the target domain. Additionally, a plausible explanation for the underlying mechanism for performance issues of unsupervised domain adaptation in case of dataset imbalance has been proposed. However, further research on the influential factors is needed as the analysis of CAMs only offered limited insights.

It was observed that a combination of domain adaptation and augmentation can address their respective shortcomings. Further investigation is needed, particularly regarding the impact of this combination on CAMs.

In the context of CAM quantification, no connection between performance and wIoU in domain adapted models and the augmented models could be observed. Therefore, models exhibiting high performance do not consistently prioritize skin lesions. Moreover, the plausibility of Grad-CAM-Elementwise activation does not indicate whether domain adaptation or augmentation led to significant performance improvements.

In contrast to the inter-domain model it appears that with augmented or DANN models, CAMs of correctly classified images are not necessarily more plausible than those of incorrectly classified images. This implies that the Grad-CAM-based explainability methods have limited potential for establishing trust in AI-based skin cancer classification in the scope of domain adapted models.

A valuable insight that extends to AI-based classification in other safety-critical medical domains is primarily Grad-CAM's inability to reliably display activation in binary classification, where one object may portray both classes. Furthermore, domain adaptation utilizes target domain dataset bias to improve performance on the target domain, which potentially leads to a model exhibiting implausible activation along with great performance in any application.

7.1. Outlook

The presented results and their conclusions offer opportunities for various extensions. As previously mentioned, augmentation and domain adaptation improved performance in different target domains. This indicates the necessity to explore the impact of a combination of augmentation and domain adaptation on performance and Grad-CAM-Elementwise activation. Another aspect worth investigating is the influence of color bias within the dataset, as identified in the HAM loc body domain. A color preprocessing of the nevus images marks a starting point for this investigation. Furthermore, finetuning hyperparameters, such as batch size, can significantly impact the resulting model's performance in this context.

The quantification of saliency maps can be extended by the incorporation of additional quantification metrics. Boggust et al. [26] have proposed Ground Truth Coverage (GTC) and Saliency Coverage (SC). GTC reaches its maximum when the ground truth is entirely encompassed by the saliency, while SC maximizes when the ground truth covers the entire saliency. The introduction of weighted variants of these metrics has the potential to provide additional insights into the quantification of activation patterns.

Lee et al. [68] successfully directed the attention of the model towards the skin lesion through alterations in network architecture. However, adjustments to the training process could also contribute to the model's ability to produce plausible activation of domain adapted models on dermoscopic images. In the context of domain adaptation, a notable work by Zunino et al. [24] demonstrated the effectiveness of their Explainable AI (XAI) training strategy in enhancing generalization, while preserving performance on the source domain. They utilized binary annotation maps to determine which image regions contain domain invariant features, resulting in reinforced activations in the annotated area. In the context of skin cancer classification, this annotated area corresponds to a skin lesion segment. For this purpose, the previously described BCDU-Net can be employed. It would be interesting to explore how the XAI training strategy influences Grad-CAM-Elementwise activations in domain shift scenarios.

As previously mentioned, Nunnari et al. [69] have also investigated the overlap between Grad-CAM saliency maps and skin lesion segments. ResNets low feature map resolution posed various challenges and presumably limits the application of Grad-CAM in medical applications. To enhance the plausibility of saliency maps, their approach is worth pursuing within the scope of domain adaptation. Given that VGG16 does not achieve state-of-the-art performance in skin cancer classification, it is necessary to explore alternative CNN architectures for generating high-resolution CAMs that effectively capture high-level features.

List of Literature

[1] H. Kutzner *et al.*, "Overdiagnosis of melanoma – causes, consequences and solutions," *JDDG J. Dtsch. Dermatol. Ges.*, vol. 18, no. 11, pp. 1236–1243, Nov. 2020, doi: 10.1111/ddg.14233.

[2] H. Lorentzen, K. Weismann, C. S. Petersen, F. G. Larsen, L. Secher, and V. Skødt, "Clinical and Dermatoscopic Diagnosis of Malignant Melanoma: Assessed by Expert and Non-expert Groups," *Acta Derm. Venereol.*, vol. 79, no. 4, pp. 301–304, Jun. 1999, doi: 10.1080/000155599750010715.

[3] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin Cancer Classification With Deep Learning: A Systematic Review," *Front. Oncol.*, vol. 12, 2022, doi: 10.3389/fonc.2022.893972.

[4] L. Sacchetto *et al.*, "Trends in incidence of thick, thin and in situ melanoma in Europe," *Eur. J. Cancer*, vol. 92, pp. 108–118, Mar. 2018, doi: 10.1016/j.ejca.2017.12.024.

[5] G. Pellacani and S. Seidenari, "Comparison between morphological parameters in pigmented skin lesion images acquired by means of epiluminescence surface microscopy and polarized-light videomicroscopy," *Clin. Dermatol.*, vol. 20, no. 3, pp. 222–227, 2002, doi: https://doi.org/10.1016/S0738-081X(02)00231-6.

[6] A. I. Oloruntoba *et al.*, "Assessing the Generalizability of Deep Learning Models Trained on Standardized and Nonstandardized Images and Their Performance Against Teledermatologists: Retrospective Comparative Study," *JMIR Dermatol.*, vol. 5, no. 3, Sep. 2022, doi: 10.2196/35150.

[7] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[8] "ISIC Challenge." Accessed: Aug. 17, 2023. [Online]. Available: https://challenge.isic-archive.com/

[9] F. Perez, S. Avila, and E. Valle, "Solo or Ensemble? Choosing a CNN Architecture for Melanoma Classification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 2775–2783. doi: 10.1109/CVPRW.2019.00336.

[10] K. Fogelberg, S. Chamarthi, R. C. Maron, J. Niebling, and T. J. Brinker, "Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation.," *New Biotechnol.*, vol. 76, pp. 106–117, Sep. 2023, doi: 10.1016/j.nbt.2023.04.006.

[11] J. V. Tembhurne, N. Hebbar, H. Y. Patil, and T. Diwan, "Skin cancer detection using ensemble of machine learning and deep learning techniques," *Multimed. Tools Appl.*, vol. 82, no. 18, pp. 27501–27524, Jul. 2023, doi: 10.1007/s11042-023-14697-3.

[12] C. Sunarya, J. Siswanto, G. Cam, and F. Kurniadi, "Skin Cancer Classification using Delaunay Triangulation and Graph Convolutional Network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, Jan. 2023, doi: 10.14569/IJACSA.2023.0140685.

[13] A. Bissoto, E. Valle, and S. Avila, "GAN-based data augmentation and anonymization for skin-lesion analysis: A critical review," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Nashville, TN, USA, pp. 1847–1856, 2021. doi: 10.1109/CVPRW53098.2021.00204.

[14] H. L. Gururaj, N. Manju, A. Nagarjun, V. N. M. Aradhya, and F. Flammini, "DeepSkin: A Deep Learning Approach for Skin Cancer Classification," *IEEE Access*, vol. 11, pp. 50205–50214, 2023, doi: 10.1109/ACCESS.2023.3274848.

[15] S. Mishra, H. Imaizumi, and T. Yamasaki, "Interpreting Fine-Grained Dermatological Classification by Deep Learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 2729–2737. doi: 10.1109/CVPRW.2019.00331.

[16] M. S. Akter, H. Shahriar, S. Sneha, and A. Cuzzocrea, "Multi-class Skin Cancer Classification Architecture Based on Deep Convolutional Neural Network," in *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, 2022, pp. 5404–5413. doi: 10.1109/BigData55660.2022.10020302.

[17] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM Multimedia*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:5219826

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *CoRR*, vol. abs/1312.6034, 2013, [Online]. Available: https://api.semanticscholar.org/CorpusID:1450294

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[20] R. L. Draelos and L. Carin, "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:244478775

[21] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," 2017, doi: 10.48550/ARXIV.1710.11063.

[22] S. Srinivas and F. Fleuret, "Full-Gradient Representation for Neural Network Visualization," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[23] C. Yang, A. Rangarajan, and S. Ranka, "Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification.," *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2018, pp. 1571–1580, 2018.

[24] A. Zunino *et al.*, "Explainable Deep Classification Models for Domain Generalization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 3227–3236. doi: 10.1109/CVPRW53098.2021.00361.

[25] K. Morrison, A. Mehra, and A. Perer, "Shared Interest…Sometimes: Understanding the Alignment between Human Perception, Vision Architectures, and Saliency Map Techniques," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vancouver, Canada, 2023, pp. 3776–3781. doi: 10.1109/CVPRW59228.2023.00391.

[26] A. Boggust, B. Hoover, A. Satyanarayan, and H. Strobelt, "Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, in CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. doi: 10.1145/3491102.3501965.

[27] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-Directional ConvLSTM U-Net with Densley Connected Convolutions," in *2019 IEEE/CVF International*

Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 406–415. doi: 10.1109/ICCVW.2019.00052.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[29] "InfluenceMap ResNet." Accessed: Sep. 29, 2023. [Online]. Available: https://influencemap.cmlab.dev/submit/?id=WC0Z-3Wv0Bli.ZGVlcCByZXNpZHVhbCBsZWFybmluZyBmb3IgaW1hZ2UgcmVjb2duaXRpb24&tab=0

[30] N. Gouda and J. Amudha, "Skin Cancer Classification using ResNet," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India: IEEE, Oct. 2020, pp. 536–541. doi: 10.1109/ICCCA49541.2020.9250855.

[31] A. Mehra, A. Bhati, A. Kumar, and R. Malhotra, "Skin Cancer Classification Through Transfer Learning Using ResNet-50," in *Emerging Technologies in Data Mining and Information Security*, vol. 1300, A. E. Hassanien, S. Bhattacharyya, S. Chakrabati, A. Bhattacharya, and S. Dutta, Eds., in Advances in Intelligent Systems and Computing, vol. 1300. , Singapore: Springer Nature Singapore, 2021, pp. 55–62. doi: 10.1007/978-981-33-4367-2_6.

[32] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 06, no. 02, pp. 107–116, Apr. 1998, doi: 10.1142/S0218488598000094.

[33] P. Ruiz, "Understanding and visualizing ResNets," Medium. Accessed: Aug. 25, 2023. [Online]. Available: https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8

[34] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," p. 244, 2016, doi: 10.48550/ARXIV.1603.07285.

[35] J. Gildenblat, "Advanced AI explainability for PyTorch." Aug. 16, 2023. Accessed: Aug. 16, 2023. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

[36] "Data augmentation," *Wikipedia*. Jul. 23, 2023. Accessed: Aug. 16, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Data_augmentation&oldid=1166738921

[37] V. Lyashenko, "Data Augmentation in Python: Everything You Need to Know," neptune.ai. Accessed: Aug. 16, 2023. [Online]. Available: https://neptune.ai/blog/data-augmentation-in-python

[38] P. M. Kazaj, M. Koosheshi, A. Shahedi, and A. Vafaei Sadr, "U-Net-based Models for Skin Lesion Segmentation: More Attention and Augmentation," *ArXiv E-Prints*, p. arXiv:2210.16399, Oct. 2022, doi: 10.48550/arXiv.2210.16399.

[39] M. Orbes-Arteaga *et al.*, "Augmentation based unsupervised domain adaptation." 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247058857

[40] S. Moon, "How data augmentation affects machine learning," Datahunt - Quality Data with Quality AI. Accessed: Aug. 17, 2023. [Online]. Available: https://www.thedatahunt.com/en-insight/how-data-augmentation-impacts-machine-learning

[41] M. J. Cooper, P. Jain, and H. S. Sidhu, "Systems and Methods for Training Machine Models with Augmented Data," Apr. 16, 2020 Accessed: Aug. 17, 2023. [Online]. Available: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2020077117

[42] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A Brief Review of Domain Adaptation," in *Advances in Data Science and Information Engineering*, R. Stahlbock, G.

M. Weiss, M. Abou-Nasr, C.-Y. Yang, H. R. Arabnia, and L. Deligiannidis, Eds., Cham: Springer International Publishing, 2021, pp. 877–894.
[43] Y. Ganin *et al.*, "Domain-Adversarial Training of Neural Networks," in *Domain Adaptation in Computer Vision Applications*, G. Csurka, Ed., Cham: Springer International Publishing, 2017, pp. 189–209. doi: 10.1007/978-3-319-58347-1_10.
[44] "Domain Adaptation in Computer Vision: Everything You Need to Know." Accessed: Aug. 17, 2023. [Online]. Available: https://www.v7labs.com/blog/domain-adaptation-guide, https://www.v7labs.com/blog/domain-adaptation-guide
[45] "InfluenceMap DANN." Accessed: Aug. 17, 2023. [Online]. Available: https://influencemap.cmlab.dev/submit/?id=VnLjPvWkbQxLWys55iWv29az.ZG9tYWluIGFkdmVyc2FyaWFsIHRyYWluaW5nIG9mIG5ldXJhbCBuZXR3b3Jrcw&tab=3
[46] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241.
[47] "InfluenceMap U-Net." Accessed: Aug. 17, 2023. [Online]. Available: https://influencemap.cmlab.dev/submit/?id=VxUO20Wv936_W_pjpB.dSBuZXQgY29udm9sdXRpb25hbCBuZXR3b3JrcyBmb3IgYmlvbWVkaWNhbCBpbWFnZSBzZWdtZW50YXRpb24&tab=0
[48] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.
[49] "Jaccard index," *Wikipedia*. Aug. 12, 2023. Accessed: Aug. 17, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=1169994638
[50] "ISIC - 2019." Accessed: Aug. 19, 2023. [Online]. Available: https://www.kaggle.com/datasets/d41aa21f043061d22da21d33a0dad05b36a7ff8a340a033899aa69329cde8767
[51] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, p. 102305, 2022, doi: https://doi.org/10.1016/j.media.2021.102305.
[52] M. Combalia *et al.*, "BCN20000: Dermoscopic Lesions in the Wild," *ArXiv*, vol. abs/1908.02288, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:199472593
[53] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Aug. 2018, doi: 10.1038/sdata.2018.161.
[54] "Machine Learning Glossary," Google for Developers. Accessed: Aug. 24, 2023. [Online]. Available: https://developers.google.com/machine-learning/glossary?hl=de
[55] "RandomCrop — Torchvision main documentation." Accessed: Aug. 24, 2023. [Online]. Available: https://pytorch.org/vision/main/generated/torchvision.transforms.RandomCrop.html
[56] "RandomHorizontalFlip — Torchvision main documentation." Accessed: Aug. 24, 2023. [Online]. Available: http://pytorch.org/vision/master/generated/torchvision.transforms.RandomHorizontalFlip.html
[57] "RandomVerticalFlip — Torchvision main documentation." Accessed: Aug. 24, 2023. [Online]. Available: http://pytorch.org/vision/master/generated/torchvision.transforms.RandomVerticalFlip.html

[58] "Resize — Torchvision main documentation." Accessed: Sep. 21, 2023. [Online]. Available: https://pytorch.org/vision/main/generated/torchvision.transforms.functional.resize.html
[59] "Normalize — Torchvision main documentation." Accessed: Aug. 24, 2023. [Online]. Available: http://pytorch.org/vision/main/generated/torchvision.transforms.Normalize.html
[60] "Torchvision Github," GitHub. Accessed: Aug. 24, 2023. [Online]. Available: https://github.com/pytorch/vision/blob/c187c2b12d86c3909e59a40dbe49555d85b98703/torchvision/models/resnet.py
[61] J. Yang, "DANN-pytorch/DANN.ipynb at master · Yangyangii/DANN-pytorch," GitHub. Accessed: Oct. 07, 2023. [Online]. Available: https://github.com/Yangyangii/DANN-pytorch/blob/master/DANN.ipynb
[62] R. Azad, "Bi-Directional ConvLSTM U-Net with Densely Connected Convolutions." Oct. 05, 2023. Accessed: Oct. 07, 2023. [Online]. Available: https://github.com/rezazad68/BCDU-Net
[63] L. Zhang, "Prevent Overfitting," Department of Computer Science, University of Toronto. Accessed: Oct. 01, 2023. [Online]. Available: https://www.cs.toronto.edu/~lczhang/360/lec/w05/overfit.html
[64] S. Chamarthi, K. Fogelberg, R. C. Maron, T. J. Brinker, and J. Niebling, "Mitigating the Influence of Domain Shift in Skin Lesion Classification: A Benchmark Study of Unsupervised Domain Adaptation Methods on Dermoscopic Images," 2023, doi: 10.48550/ARXIV.2310.03432.
[65] S. Tan, X. Peng, and K. Saenko, "Class-Imbalanced Domain Adaptation: An Empirical Odyssey," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds., Cham: Springer International Publishing, 2020, pp. 585–602.
[66] J. Gachon *et al.*, "First Prospective Study of the Recognition Process of Melanoma in Dermatological Practice," *Arch. Dermatol.*, vol. 141, no. 4, pp. 434–438, Apr. 2005, doi: 10.1001/archderm.141.4.434.
[67] L. Gamage, U. Isuranga, S. De Silva, and D. Meedeniya, "Melanoma Skin Cancer Classification with Explainability," in *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, 2023, pp. 30–35. doi: 10.1109/ICARC57651.2023.10145622.
[68] D. Lee, S.-H. Lee, and J.-H. Jung, "The effects of topological features on convolutional neural networks—an explanatory analysis via Grad-CAM," *Mach. Learn. Sci. Technol.*, vol. 4, no. 3, p. 035019, Sep. 2023, doi: 10.1088/2632-2153/ace6f3.
[69] F. Nunnari, M. A. Kadir, and D. Sonntag, "On the Overlap Between Grad-CAM Saliency Maps and Explainable Visual Features in Skin Cancer Images," in *Machine Learning and Knowledge Extraction*, vol. 12844, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., in Lecture Notes in Computer Science, vol. 12844. , Cham: Springer International Publishing, 2021, pp. 241–253. doi: 10.1007/978-3-030-84060-0_16.

Declaration of Authorship

Herewith I declare that I prepared this thesis on my own, that I did not use any other sources and resources than those that are specified, that all arguments and ideas that were literally or analogously taken from other sources are sufficiently identified, and that the thesis in identical or similar form has not been used as part of an earlier course achievement or examination procedure.

Jena, 19.10.2023