Automated malignant melanoma detection using supervised contrastive learning

Cassandra Durr

Research assignment presented in the partial fulfilment
of the requirement for the degree of
MSc Machine Learning and Artificial Intelligence
at Stellenbosch University

Supervisor: Dr Johannes Coetzer

ABSTRACT

Modern computer-aided detection methodologies have proven to be highly successful in the classification of melanoma, an aggressive variant of skin cancer, from dermoscopic imagery. This research seeks to evaluate the effectiveness of the recently introduced supervised contrastive learning training regime in identifying melanoma. The regime is evaluated through comparison to a baseline approach utilising binary cross-entropy loss. A secondary objective of the study is to compare the melanoma detection performance of a vision transformer to two popular convolutional neural network (CNN) architectures, ResNet50 and InceptionV3. Contrary to prevailing literature, the study's findings show that supervised contrastive learning does not consistently surpass the baseline approach, and vision transformers do not invariably demonstrate superior performance over CNN architectures. These findings necessitate further validation through means of hyperparameter tuning and repeated experimentation, given the presence of sufficient computational resources. This study outlines the challenges associated with melanoma detection research and provides a number of recommendations to improve both the performance and equitability of the proposed computer-aided detection model.

Key words:
Melanoma detection; supervised contrastive learning; computer-aided detection

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Melanoma is a highly aggressive form of skin cancer that accounts for the majority of skin cancer fatalities (American Cancer Society, 2017). Early detection of melanoma is critical for positive prognoses, as melanoma is most likely to be cured in its early stages (American Cancer Society, 2019). However, melanomas are often inconsistent in appearance, which makes them difficult to detect, especially in the early stages when their presentation may not be markedly different from benign lesions or moles (National Cancer Institute, 2011). Inconsistencies in opinion between healthcare professionals visually assessing lesions also adds to the challenges associated with melanoma detection. Studies have shown that the rules-based approaches and heuristics used to diagnose melanomas in industry do not achieve consistent and reproducible results (Masood & Ali Al-Jumaily, 2013).

In response to the aforementioned challenges, computer-aided detection (CAD) methodologies have emerged to automatically determine whether a lesion is malignant or benign. CAD systems have shown to outperform primary care physicians and perform at least as well as expert dermatologists in terms of ROC AUC (Phillips et al., 2020). The historical success of automated melanoma detection algorithms suggests that such algorithms could be used in practice to assist medical practitioners in their diagnostic procedures.

This study outlines a comprehensive approach to determine an appropriate training regime and neural network architecture tailored to the challenge of melanoma detection. This study employs both supervised contrastive learning and a baseline approach utilising binary cross-entropy loss to classify dermoscopic images as either benign or malignant. In the pursuit of optimising melanoma detection performance, this study evaluates three encoder architectures: the recently introduced vision transformer, and two widely-utilised CNNs - ResNet50 and InceptionV3.

Historically, contrastive learning has been used as a self-supervised representation learning method. Self-supervised contrastive learning produces pseudo-labels from unlabelled data and contrasts image samples against each other to learn features that are similar between classes and those that set classes apart from one another (Kundu, 2023). In 2021, researchers extended this framework to the fully-supervised setting and outperformed previous state-of-the-art classification results (Khosla et al., 2020).

This study aims to contribute to the development of reliable and efficient methods for the detection of melanoma by overcoming the limitations of traditional detection systems and exploring the performance of supervised contrastive learning in this setting.

1.1 DATA

The data utilised in this study is the HAM10000 dataset made available by Tschandl (2018). The HAM10000 dataset consists of 10,015 training images and 1,511 test images. The ground-truth test dataset labels were only made available in 2023 as the dataset formed part of an online melanoma detection challenge hosted by the International Skin Imaging Collaboration (ISIC).

Each image in the training and test datasets corresponds to a record of patient metadata. The data available per lesion includes the sex and age of the individual, the localisation of the lesion on the patient's body and the method of lesion diagnosis. Histopathology confirms the diagnosis for the majority of the lesion images ($> 50\%$), while expert consensus or in-vivo confocal microscopy is used to determine the remaining diagnoses. The tabular features per lesion were not utilised in this study; however, Section 5.3 describes how metadata could be incorporated in future iterations of this project.

While the dataset comprises labeled data related to seven distinct skin conditions, the objective of this study is melanoma detection. The dataset is therefore adapted for binary classification in this study.

1.2 PROBLEM STATEMENT

The primary focus of this study is to determine whether the supervised contrastive loss formulated by Khosla et al. (2020) offers an improvement over baseline classification models trained with binary cross-entropy loss. In the aforementioned paper, the authors compare the performance of their proposed supervised contrastive loss to the popular cross-entropy loss. This research, however, deals with a binary classification problem which renders binary cross-entropy loss a more suitable baseline. The study therefore aims to answer the following research question:

> *Does supervised contrastive learning outperform traditional image classification models in detecting melanoma?*

The secondary objective of this study is to compare the performance of a vision transformer against standard convolutional neural network (CNN) architectures under a fixed computational budget. Vision transformers have recently dominated computer vision literature due to their ability to better capture long-range spatial dependencies in comparison to CNNs. Therefore, a secondary objective of this study is to answer the question:

> *Do vision transformers yield superior classification performance over CNNs for the task of melanoma detection?*

1.3 CHAPTER PLAN

In Chapter 2, the foundations for supervised contrastive learning are presented, including an explanation of the supervised contrastive loss function known as 'SupCon', formulated by Khosla et al. (2020). Additionally, the chapter gives an overview of various approaches proposed in academic literature to address the melanoma detection issue.

The methodology employed in this research project is described in detail in Chapter 3. This chapter explains the modelling decisions taken in light of several challenges such as restricted computational resources and class imbalance in the HAM10000 dataset. Furthermore, the chapter describes the modifications made in comparison to the initial 'Supervised Contrastive Learning' paper authored by Khosla et al. (2020).

Chapter 4 centers on the results of this study, including a discussion on how these findings align with initial expectations and existing literature.

In Chapter 5, the limitations of this study is discussed, and improvements are proposed for further extensions of this work.

Lastly, Chapter 6 provides a conclusion, summarising the main findings and contributions of this project.

CHAPTER 2

LITERATURE REVIEW

This chapter establishes the foundation for this research project by delving into the fundamentals of supervised contrastive learning. Subsequently, the existing architectures used to address the melanoma detection problem will be explored.

2.1 SUPERVISED CONTRASTIVE LEARNING

Contrastive learning, which constitutes a representation learning method, aims to automatically discover feature patterns in a dataset and encode the features into numerical representations of lower dimension than the original data, where similar images have similar encodings and dissimilar images have distinct, dissimilar encodings. Supervised contrastive learning seeks to teach an encoder how to distinguish between images belonging to the same class and images belonging to different classes. This process allows the model to learn a discriminative embedding space by pulling together the encodings of images from the same class while simultaneously separating the encodings of images from different classes.

The supervised contrastive learning framework follows a staged approach where augmented images are passed through an encoder and projection head, both trained with a contrastive loss function. Subsequently, the projection head is discarded, and a multi-layer perceptron is trained using a classification loss function, with the frozen encoder as its input. The projection head is discarded due to research findings indicating that projected features obtained from contrastive loss training tend to exhibit lower generalisation performance compared to the features extracted directly from the encoder (Gupta et al., 2022).

In their paper 'Supervised Contrastive Learning,' Khosla et al. (2020) demonstrate that their proposed training regime, which includes supervised contrastive loss, achieves superior classification accuracy compared to cross-entropy loss on the ImageNet dataset. Moreover, this approach proves to be less sensitive to a variety of hyperparameters. However, it is essential to note that achieving such impressive performance requires significantly large batch sizes. The authors provide top-1 accuracy results across batch sizes ranging from 512 to a maximum of 6,144, on architectures ranging from 20 million to 90 million parameters. In Section 2.1.1, it is shown that the images processed in a single batch is double the original batch size due to image augmentation, therefore the effective batch sizes utilised in the study range from 1,024 to 12,288. Whilst, Khosla et al. (2020) showcase state-of-the-art classification performance on ImageNet, the batch sizes utilised in their study demand substantial memory and computational resources.

The effectiveness of contrastive loss relies on the number of positive and negative instances per

anchor image. A positive instance refers to an image with the same class label as the anchor from a batch, while a negative instance is an image with a different class label from a batch. As the number of in-batch positives and negatives per anchor image increase, the contrastive power of the loss function strengthens. To fully exploit contrastive loss and encourage the model to learn meaningful representations, it is critical to have a sufficiently large batch size.

As melanoma detection involves only two classes, compared to ImageNet's 1000 classes, the requirement for exceedingly large batch sizes may be reduced. As mentioned previously, the quality of the learnt embeddings trained using contrastive loss is greatly dependent on the number of positive and negative instances per anchor image. Applying contrastive learning to the ImageNet classification task requires the model to develop an embedding space in which hundreds of classes must be distinctly differentiated. To achieve this objective, each anchor image needs to be contrasted against images from a wide array of classes, which in turn requires substantially large batch sizes. Conversely, in the context of binary classification tasks, the model's objective is limited to distinguishing between two classes. Hence, the requirement for a substantial batch size is lessened as the batches do not need to include images from a diverse array of classes. The binary nature of the melanoma detection problem allows the contrastive learning model to focus on differentiating between images across only two classes, and as a result may learn meaningful and distinct embeddings with reduced batch sizes.

2.1.1 SupCon Loss

In this subsection, a more thorough exploration is conducted into the loss function employed in the paper titled 'Supervised Contrastive Learning' by Khosla et al. (2020), as it plays a crucial role in the methodology applied within this research study.

The authors augment each image in a batch twice, therefore two views of an image exist in a batch of training. The augmentations are considered positive instances and share a class label. If a batch of data includes N original images, the 'multi-viewed' batch consists of $2N$ images. Given this notation, the contrastive loss function presented by Khosla et al. (2020), referred to as the 'SupCon' loss, has the form

$$\begin{aligned} L &= \sum_{i \in I} L_i \\ &= \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \ln \left(\frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \right) \end{aligned} \tag{2.1}$$

where:

- $I = \{1, 2, ..., 2N\}$ is the set of indices of the images in a batch after augmentation.
- $\tau \in \mathbb{R}^+$ represents a hyperparameter, called temperature.
- $P(i)$ is the subset of indices from I, corresponding to images in the batch which share the same label as image i, excluding index i. Therefore, $|P(i)|$ represents the number of elements belonging to set $P(i)$, or equivalently, the cardinality of $P(i)$.
- $A(i)$ denotes the subset of I representing indices in a batch corresponding to images with a different class label to that of image i.
- Lastly, $\mathbf{z}_j$ represents the encoding, or latent representation, for the image corresponding to index j. This encoding is the output of the projection head, as depicted in Figure 3.1, and not the output of the encoder. The projection head operates on the encoder's output and projects it onto a low-dimensional latent space.

The loss L is computed as an aggregation of losses per anchor image within a batch. Since the batching process is stochastic, each anchor image may have a varying number of positive and negative instances per epoch. Nevertheless, the process of augmenting each image twice ensures that every anchor image is guaranteed to have at least one positive instance. Given the large number of classes in ImageNet, each batch will contain a mix of classes, assuming a sufficiently large batch size. Therefore, in any given batch each anchor will have many corresponding negatives. Since each anchor is guaranteed to have at least one positive and negative instance, the loss per anchor can be computed, and the contrastive learning objective remains well-defined.

The proposed loss function has been designed to guide the model's convergence towards optimal solutions by leveraging the pairwise similarities between each anchor image and their corresponding positive and negative instances. The dot product, used in the loss function, can be thought of as a metric of similarity due to its relationship with cosine similarity. The equation for cosine similarity between two numerical vectors, $\mathbf{u}$ and $\mathbf{v}$, is

$$\text{Cosine similarity}(\mathbf{u}, \mathbf{v}) = \cos(\theta) \tag{2.2}$$

$$= \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{2.3}$$

where θ represents the angle between the vectors in Euclidean space. The equation for cosine similarity can be re-organised so that $\mathbf{u} \cdot \mathbf{v} = \cos(\theta)\|\mathbf{u}\|\|\mathbf{v}\|$. Consequently, when vectors $\mathbf{u}$ and $\mathbf{v}$ are aligned, their dot product will be large, and when two vectors are greatly dissimilar, their dot product will be small.

Therefore, the dot product between the latent representation of the anchor i and the latent repre-

sentation of positive instance p, denoted $\mathbf{z}_i \cdot \mathbf{z}_p$, should be large when the encodings are well aligned. Moreover, the summation in the denominator of the loss function, $\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)$, should comprise small terms if the encodings of anchor and negative pairs lie far apart in the embedding space. A well-trained model should then achieve a large value for

$$\ln \left(\frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \right) \tag{2.4}$$

since $\ln(\cdot)$ and $\exp(\cdot)$ are monotonic transformations. Because the term

$$\frac{1}{|P(i)|} \sum_{p \in P(i)} \ln \left(\frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \right) \tag{2.5}$$

is negated in the presented loss equation, the loss per anchor i for a well-trained system will be small, and the learnt encodings for images will be suitably defined.

2.2 RELATED WORK

Determining whether an image of a skin lesion is a malignant melanoma is a well-studied problem in machine learning. In this section, the different methodologies that have been used to solve the problem will be outlined.

2.2.1 Hand-Crafted Diagnostic Features

A well-known diagnostic tool for diagnosing melanoma is the 'ABCD' tool (American Academy of Dermatology Association, 2023) which is described below.

- **Asymmetry** - asymmetry in lesions indicate melanoma.
- **Borders** - the borders of melanomas are typically irregular and blurry.
- **Colour** - melanomas are typically multicoloured.
- **Diameter** - melanomas generally have a diameter exceeding 6 mm.

For clarity, these criteria are visually depicted in Figure 2.1.

Historically, many researchers have coded these hand-crafted traits as input features to a classification model. Some papers use these hand-crafted features in isolation to classify melanoma, such as the papers by Cueva et al. (2017), Jain et al. (2015) and Thanh et al. (2020). Other papers use the hand-crafted features in combination with features learnt from deep learning models such as the papers by Almaraz-Damian et al. (2020) and Ichim & Popescu (2020). Models using 'ABCD' fea-

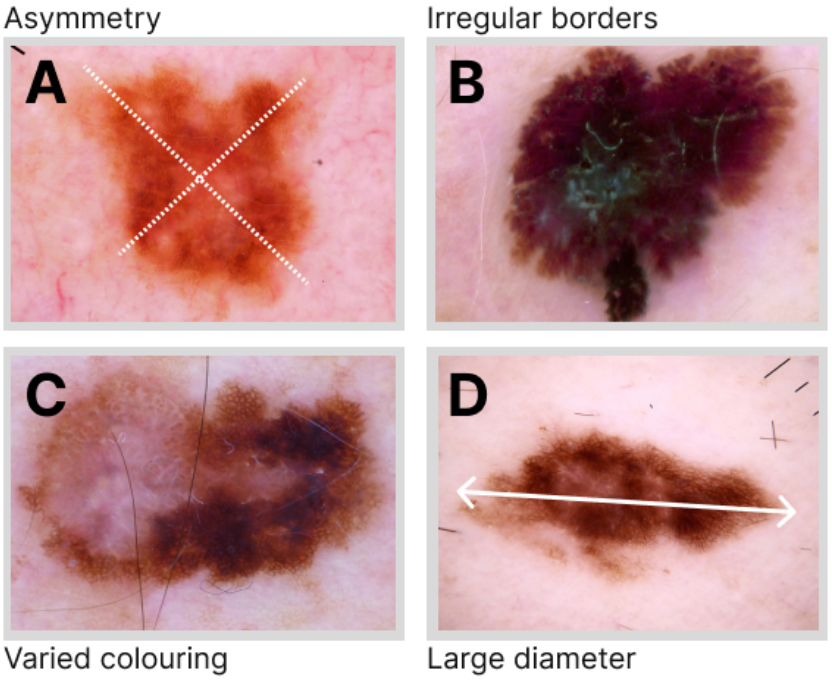

Figure 2.1: A visual representation of the melanoma markers as per the ABCD diagnostic rule. Each panel within this figure illustrates a different diagnostic criterion. Images are sourced from the HAM10000 training data.

tures, and other hand-crafted features perceived to be of importance, typically extract the feature information using image segmentation and edge detection.

The issue with using hand-crafted model inputs is that the features are based on specific rules or heuristics, and may not be flexible enough to fully capture the complex patterns present in skin lesion datasets. Machine learning models can evolve as they are exposed to more data, whereas hand-crafted features remain static. Due to the rigidity of hand-crafted features and the fact that most modern day CAD algorithms use only learnt features, the remainder of this analysis will focus on algorithms using learnt features.

2.2.2 Ensembles

The key idea behind ensemble learning in this context is to boost classification accuracy by assembling multiple classification models which may have different architectures. Due to the stochasticity inherent to deep learning models, individual models will learn different traits from a dataset and will make different mistakes. Therefore, by combining the knowledge learnt from a host of individual models, predictions from an ensemble model will likely achieve higher accuracy and will be more robust since weaknesses in any individual model are compensated for by the remaining models.

Many melanoma detection papers use ensembling, such as those by Mahbod et al. (2019), Mahbod et al. (2021), Shehzad et al. (2023) and many others. In addition to this, all of the prize winners of the 2020 Kaggle SIIM-ISIC Melanoma Classification competition use ensembling (Kaggle, 2020). While ensembling undoubtedly enhances the performance of melanoma detection models, it comes at the cost of long training times and significant computational expense. Additionally, ensembled models tend to be complex and challenging to interpret. Due to the restricted computational budget of this research project, the decision was made not to pursue ensembling.

2.2.3 Convolutional Neural Networks

Convolutional neural networks have been widely used in the field of computer vision and form the basis of most published literature on automatic melanoma detection. A recent survey study by Popescu et al. (2022) listed papers that make use of popular neural network architectures in detecting melanoma. The largest subset of architectures listed in the survey study consist of CNNs, and ensembles thereof. The CNN architectures which have been applied to the melanoma detection problem include ResNet, Inception/GoogLeNet, U-Net, DenseNet, AlexNet, Xception, EfficientNet, VGG, NASNet, and MobileNet. This extensive list illustrates that many CNN architectures have been proposed for the melanoma detection problem, and many show impressive performance.

In this project, the Inception and ResNet models are used as encoders. These CNN encoders were chosen because of their popularity in literature as well as their proven success in learning meaningful representations for melanoma detection.

2.2.4 Vision Transformers

Vision transformers (ViTs) represent a recently introduced transformer architecture specifically designed for learning image features and facilitating image classification tasks. This architecture was first introduced by Dosovitskiy et al. (2020) in October 2020 and has since been cited almost 20,000 times to date which illustrates how quickly this architecture has been adopted.

Unlike traditional CNN architectures, vision transformers rely on a self-attention mechanism that allows the model to capture long-range spatial dependencies in images. CNNs lack this ability to understand long-range spatial relationships because their architecture is based on local receptive fields and convolutional operations. CNNs are therefore unable to consider the relationship between distant parts of an image. Whether understanding long-range spatial dependencies are critical to melanoma detection is unclear, although papers which have compared the performance of ViTs to CNNs in detecting melanoma have shown favourable ViT performance. Papers which performed melanoma detection using vision transformers, and report superior performance to architectures using CNNs, include Aladhadh et al. (2022), Xin et al. (2022), and Cirrincione et al. (2023).

The 'ABCD' diagnostic tool, as described in Section 2.2.1, suggests that lesion asymmetry is an important predictor for melanoma. A significant portion of the skin lesion images from publicly available datasets are heavily magnified, and standard pre-processing often involves center cropping, which further amplifies the input. Consequently, opposite sides of most skin lesions following pre-processing will lie on opposite sides of the input images. This poses a challenge for CNNs which have a limited sense of long-range spatial patterns, since the parts of the image which will suggest asymmetry will lie far apart in the input. Similarly, the 'ABCD' framework suggests that having a great colour diversity in a lesion is a risk factor for melanoma. If a model lacks understanding of long-range spatial relationships in an image, it may fail to detect diverse colour distributions across the full image. These are just two examples that could be used to explain why vision transformers tend to outperform CNN architectures in melanoma detection.

In this research, the performance of a vision transformer encoder is compared to that of two CNNs in order to understand whether having better awareness of long-range spatial patterns improves melanoma detection performance.

2.2.5 Contrastive Learning

This section aims to explore existing models which have used a contrastive loss function within the training of a melanoma detection model.

The paper by Xin et al. (2022) uses a Siamese network architecture with a vision transformer as the encoder. The encodings of lesion images are fed into a contrastive loss function which is trained alongside the classic cross-entropy function in a multi-task fashion. The training regime followed in this research is staged instead of multi-tasked, and mirrors the training regime from the original supervised contrastive loss paper by Khosla et al. (2020). However, the data pre-processing followed in this research study is largely identical to the pre-processing employed by Xin et al. (2022), including balanced sampling to reduce the effects of class imbalance. Xin et al. (2022) find that including the contrastive objective improves the accuracy on the HAM10000 dataset (Tschandl et al., 2018) from 0.939 to 0.943.

The paper by Hsu & Tseng (2022) also makes use of supervised contrastive learning to predict melanoma in skin lesion images, but uses a proposed loss function that the authors call 'hierarchy-aware contrastive loss' in conjunction with supervised contrastive loss. The encodings that are produced using supervised contrastive loss and hierarchy-aware contrastive loss are passed through separate classifiers and the final classification result is based on an average of the two classifier results. The encoder architecture employed in the aforementioned paper constitutes a CNN, EfficientNet-B4.

A contrastive loss paper presented by Chen et al. (2022) introduce a framework similar to the one proposed in this research for melanoma classification. The authors utilise focal loss, instead of binary cross-entropy loss, to address the impact of the imbalanced dataset, while balanced sampling is proposed in this research. In addition to this, Chen et al. (2022) focus solely on the ResNet architecture for encoding, while multiple encoder options are investigated in this study.

CHAPTER 3

METHODOLOGY

The goal of this chapter is to describe the modelling choices made to train and evaluate a melanoma detection system. The proposed framework is illustrated in Figure 3.1.

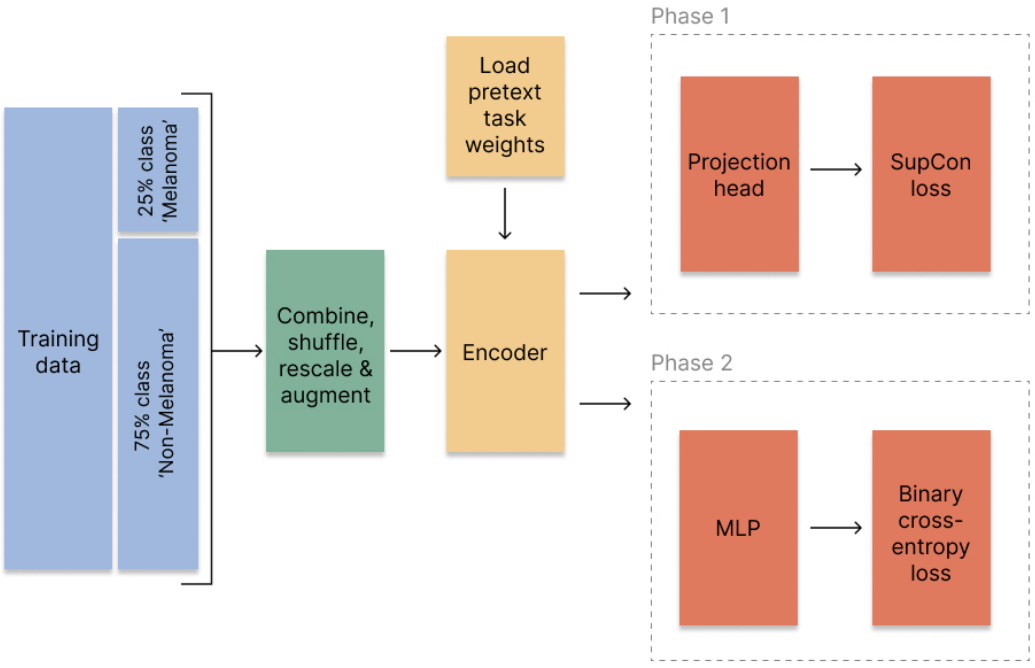

Figure 3.1: A flowchart illustrating the study's proposed training methodology for melanoma detection.

3.1 PRETEXT TRAINING

Transfer learning has revolutionised computer vision, especially since the introduction of the large-scale annotated image dataset, ImageNet by Deng et al. (2009). Transfer learning allows for the utilisation of knowledge encoded in pre-trained models which are trained on large annotated datasets. These models can be fine-tuned through training to adjust their weights to suit a slightly different domain. In modern literature, many researchers fine-tune models using pre-trained weights originally applied to the ImageNet classification task. This approach oftentimes circumvents the lengthy training periods associated with training a model from scratch, rendering computer vision more accessible for individuals with constrained computational resources. Additionally, using pre-trained weights can reduce the necessity for curating a significantly large annotated dataset for a specific task, a process that can be both costly and time-consuming.

In this study, the decision has been made not to utilise pre-trained weights. This choice is attributed to the significant dissimilarity between skin lesion images and the images found in ImageNet, or other publicly available datasets for which pre-trained weights are obtainable. The significant differences between the HAM10000 dataset and datasets like ImageNet would render the pre-trained weights unhelpful for the task of melanoma detection.

An alternative approach to using pre-trained weights involves training a network using a pretext task to familiarise the model with the specific domain, such as dermatology images in this context. To prevent overfitting, the same dataset should not be used for both pretext training and the subsequent training phase. By adopting this strategy, the model can encode features from the specific domain without becoming biased to the training data, which should enhance the model's downstream melanoma detection performance.

The Stanford University 'Diverse Dermatology Images' dataset, made available by Daneshjou et al. (2022), was used for pretext training, where the chosen pretext task was image denoising. The encoder weights from the pretext models were saved and then subsequently loaded during the melanoma detection training. A further iteration of this study may consider whether alternative pretext tasks improve the downstream melanoma classifier.

The choice of the pretext dataset holds significance. According to the journal article by Daneshjou et al. (2022), publicly available dermatology datasets exhibit a bias towards images of lesions on lighter skin tones, resulting in less proficient models when classifying skin diseases on images of individuals with darker skin tones. The HAM10000 dataset does not include features describing ethnicity or skin tone; however, visualising a subset of the data reveals that the majority of images belong to individuals with lighter skin. In contrast, the Stanford DDI dataset includes skin lesion images from individuals with a diverse range of skin tones. Exposing the model encoder to these diverse images during pretext training should improve its generalisation to individuals with skin tones that are underrepresented in the HAM10000 dataset.

A suggested model improvement is proposed in Section 5.1 that incorporates an inference mechanism assessing the skin tone of individuals in skin lesion images according to the Fitzpatrick scale. If the skin tone is identified as being underrepresented in the training data, the model generates a prediction confidence score and includes a cautionary note in the output acknowledging the potential inaccuracies due to underrepresentation. Including equitable dermatology datasets in training and accounting for the effects of underrepresentation should promote safer and more robust models.

3.2 DATA PARTITIONING

The HAM10000 dataset comprises of 10,015 training images and 1,511 test images. The allocation of the dataset within this study is shown in Table 3.1. The training data was divided into a training partition and a validation partition in such a way that the distribution of classes was preserved across the two partitions.

Table 3.1: The number of images in each data partition by class.

Data Partition	Melanoma	Non-Melanoma	Total
Training	947 (11.12%)	7,567 (88.88%)	8,514
Validation	166 (11.06%)	1,335 (88.94%)	1,501
Testing	171 (11.32%)	1,340 (88.68%)	1,511

Table 3.1 illustrates that across all data partitions, the distribution of classes is skewed heavily towards the class 'non-melanoma'.

3.3 CLASS IMBALANCE

As mentioned in Section 3.2, HAM10000 exhibits a significant class imbalance where the set of non-melanoma images make up the majority class. Class imbalance poses a challenge in classification tasks, such as melanoma detection, as classifiers tend to exhibit bias towards the majority class resulting in poor performance and generalisation.

Suppose a classifier is trained on a dataset where 95% of the training instances belong to the majority class. If, in the extreme case, this classifier predicts the majority class 100% of the time, the classifier will achieve an impressive accuracy of 95% on the training data, however *all* instances of the minority class will be misclassified. This error is critical when the task is disease detection, where detecting the minority class (melanoma) is of utmost importance for patient diagnosis and subsequent treatment decisions. A highly imbalanced dataset may lead to a high proportion of false negatives, meaning that the classifier has failed to identify melanoma in images of lesions. As a consequence, critical medical intervention and patient care may be delayed or withheld completely.

A number of approaches exist to combat class imbalance, some of which include oversampling the minority class, undersampling the majority class, and weighting the contribution of the classes to the loss function. In Section 3.5, the approach taken to mitigate the impact of class imbalance is discussed.

3.4 DEVIATION FROM SUPCON LOSS

The methodology employed in this project mirrors that outlined in the original supervised contrastive learning paper by Khosla et al. (2020), with a single deviation.

As discussed in Section 2.1.1, Khosla et al. (2020) employ a method where each image in a batch is augmented twice, and these augmentations are treated as positive pairs along with the other augmentations in the batch that share the same class label. Retaining both sets of augmentations per original image per batch effectively doubles the dataset's size and significantly increases the memory demands for training. Since supervised contrastive learning already demands significant computational resources, this additional burden would be impractical for the setup of this research project.

To address these computational restrictions, the SupCon loss function is modified in this study. In the original approach, positive instances for an anchor image include its augmentation and all other images in the batch that share the same label. The approach adopted in this study involves augmenting each image in the batch only once, and all other augmented images with the same label are regarded as positive instances for an anchor image. This adaptation maintains the essence of contrastive learning by ensuring that each augmented image contributes to multiple positive pairs, while alleviating memory constraints.

The adaptation, although computationally necessary, introduces a complication. In the original formulation, each anchor image was guaranteed to have at least one positive instance, its augmentation. However, in the new formulation, an anchor image may lack a corresponding positive instance if no image in the batch shares the same class label with the anchor. This issue is further aggravated by the dataset's severe class imbalance, as described in Section 3.3. The class distribution is skewed to the extent that, under a traditional batching strategy, there is a high probability of forming a batch entirely composed of skin lesions that are not melanomas. Consequently, in such a scenario, no negative instances will be available for any image in the batch since the batch does not represent a mix of the two available classes.

To ensure that each batch consists of a mix of images from the two classes, a custom batching strategy was employed and is discussed further in Section 3.5.

3.5 CLASS-WEIGHTED BATCHING APPROACH

To ensure well-defined losses per anchor, it is essential for each batch to contain a minimum of two melanoma images and two non-melanoma images, as discussed in Section 3.4. However, when augmenting each image only once, the traditional batching method used by Khosla et al. (2020),

where the dataset is shuffled and split into mini-batches, does not guarantee class mixing in each batch. To overcome this limitation of traditional batching, and to mitigate the effect of class imbalance, a more sophisticated sampling strategy is adopted in this study that creates batches with sufficient representation from both classes.

Initially, the possibility of using the `kerasgen` data generator developed by Bahaa (2022) was explored. This generator constructs batches with a perfect balance between the two available classes by oversampling the minority class (melanoma). Experimental results of this study demonstrated that achieving a perfect class balance led to poor generalisation performance, particularly resulting in a high false positive rate in unseen datasets. The issue arose because excessively correcting the class imbalance caused the model to expect approximately half of the inference instances to belong to the melanoma class, however the true class distribution is heavily skewed toward the non-melanoma class. Consequently, the model predicted many negative instances as positives, producing very high recall scores, but poor precision values. Consequently, the decision was made not to use this generator, and instead, a custom generator was built to allow for specifying the mix of classes.

The custom generator developed constructs batches in such a way that $x\%$ of the images in a batch belong to the class 'non-melanoma', while the remaining $(100 - x)\%$ are melanomas, where x is a specified value in the range of $(0, 100)$. This approach offers the flexibility to control the extent of correction for the class imbalance. By adjusting the value of x, the representation of melanoma and non-melanoma images in each batch can be effectively balanced, leading to a better trade-off between addressing class imbalance and maintaining the model's generalisability. In this study the value of x is chosen as 75, however x should ideally be tuned.

One limitation of the developed generator is that batches are sampled independently, which means that not all training instances are equally seen during training. Some instances may be encountered multiple times, while others may not be seen at all in certain epochs. This drawback is also present in the `kerasgen` data generator. Fortunately, the training dataset is extensive and diverse, and the model is allowed to train for many epochs, which helps mitigate the impact of uneven sampling. The ample data and extended training duration ensures that the model can encounter a wide range of examples over the course of its training.

In a future iteration of this project, the generator should be improved to include dependent batch sampling. The generator should maintain a record of previously sampled images and ensure that subsequent batches are selected from the set of images that have not been encountered before, until each image has been seen at least once. After ensuring the coverage of all images per class in the dataset, the generator can proceed to resample from the set of images already seen to maintain

the desired class mix. This approach guarantees that all training instances are seen at least once before any resampling occurs.

Through the implementation of this improvement, the training process can be optimised, and the model's capacity to learn from the entire dataset is likely to improve, potentially resulting in enhanced performance and generalisation. Nevertheless, this improvement comes with certain trade-offs. Specifically, it leads to increased memory consumption due to the need to keep a record of previously sampled images. Additionally, the process of determining the set from which to sample becomes more computationally intensive, resulting in longer training times. Considering the substantial computational restrictions of this research project, the implementation of this enhancement is not currently feasible. As a result, this study makes use of a generator that samples batches independently.

3.6 DATA AUGMENTATION

Numerous papers have attributed strong data augmentation as a key factor contributing to the impressive performance and generalisability of computer vision models. Therefore, this study utilises a number of data augmentation techniques such as incorporating random vertical and horizontal flips, introducing a small amount of random Gaussian noise, performing random rotation, and applying random zooming.

3.7 STAGED TRAINING APPROACH

Figure 3.1 illustrates the staged training process, which is divided into two phases. In the first phase, the 'SupCon' contrastive loss defined in Equation 2.1 is utilised to update the encoder and projection head weights. In the second phase, the encoder weights are frozen, and the projection head is discarded, as discussed in Section 2.1. Subsequently, the multi-layer perceptron weights, situated after the encoder, are trained using the binary cross-entropy loss function for melanoma classification. This methodology is in accordance with the approach introduced in the paper titled 'Supervised Contrastive Learning' by Khosla et al. (2020). The baseline model for comparison is constructed with the phase two architecture and is trained end-to-end using the binary cross-entropy loss function.

3.8 ADJUSTING THE PREDICTION THRESHOLDS

In this study, the cost associated with a false negative outweighs that of a false positive. A false negative entails a patient erroneously believing they are cancer-free, leading to a potentially delayed diagnosis. This delay can allow cancer advancement to such a point that effective treatment choices

may be limited. In contrast, a false positive results in a patient undergoing an unnecessary biopsy to confirm the diagnosis. While the cost associated with false positives are non-negligible, the consequences of false negatives are far more severe. Consequently, prioritising the reduction of false negatives is safer and vital for timely cancer detection.

Ideally, a classifier should exhibit both a low false positive rate and a low false negative rate. However, these metrics tend to have a trade-off relationship known as the precision-recall trade-off. If a positive instance is associated with the 'melanoma' class, precision measures the ratio of accurately identified positives among all of the model's predicted positive instances. In contrast, recall measures the fraction of the model's positive predictions in relation to all potential true positives within the dataset. The F_β score is a weighted harmonic mean of precision and recall, restricted to the range $[0, 1]$. An F_β value of 1 represents a perfect score, while 0 indicates the poorest possible performance. The formula for F_β is

$$F_\beta = \frac{(1 + \beta^2)(\text{precision} \times \text{recall})}{\beta^2 \times \text{precision} + \text{recall}} \tag{3.1}$$

where β determines the contribution of recall and precision to the score. When $\beta < 1$, precision is weighted higher than recall, when $\beta = 1$ the metrics are equally weighted and lastly, when $\beta > 1$ recall has a larger contribution than precision. In this study, recall is considered twice as important as precision and therefore β is chosen as 2 when evaluating the classifier's performance with F_β.

The output of the melanoma detection classifier is a value ranging between 0 and 1, generated by the sigmoid activation function. This value represents the probability that an image corresponds to the 'melanoma' class. Typically, if the probability exceeds 0.5, the image is classified to class 'melanoma', but this threshold can be adjusted. A higher classification threshold would potentially result in an increase in false negatives, whilst a lower threshold may result in more false positives. In this study, the threshold selection process involves maximising the $F_{\beta=2}$ score on the validation dataset, which is randomly sampled and not subjected to balanced batching. Henceforth, $F_{\beta=2}$ will be referred to as F_2. By choosing the threshold to optimise the F_2 score, the classifier's predictions are geared towards minimising false negatives, while still considering the false positive rate.

CHAPTER 4

RESULTS

This chapter presents the results of this study with a discussion interpreting the results in relation to the proposed problem statements, and to existing literature.

4.1 RESULTS

Table 4.1 presents key classification metrics on the validation dataset for baseline models and models trained using supervised contrastive learning across three encoder architectures. The metrics considered in this study include AUC (area under the receiver operating characteristic curve), accuracy, recall, and F_2. The importance of recall and F_2 to this study is described in detail in Section 3.8. The metric AUC is chosen because it measures the quality of the model's predictions whilst being agnostic to the choice of classification threshold. Accuracy is selected as a suitable metric because it acts antagonistically to recall. The classification threshold is chosen to optimise the F_2 metric which prioritises recall. When recall is elevated, a decline in accuracy is anticipated as high recall typically arises when the threshold is much lower than the default. In such a situation, many negative instances will be incorrectly classified as positive instances, causing accuracy to sharply drop due to the imbalanced distribution of classes in the data. In this research, prioritising recall is critical; however, it should not come at the cost of an excessively low accuracy score.

Additional information provided in Table 4.1 includes the number of parameters per encoder architecture, the chosen threshold for classification, and the mean score and 'best score' per architecture and batch size configuration. The 'best score' metric quantifies the number of metrics in which an encoder demonstrates superior performance compared to the corresponding encoder trained using a different method. To illustrate, one can consider the ViT architecture trained using the supervised contrastive learning methodology, and a batch size of 16. The 'best score' for this model is the number of metrics in which this model outperformed the baseline ViT architecture utilising the same batch size. Additionally, the mean score per batch size and encoder architecture represents the mean of the recorded metrics on the validation dataset.

Across all experiments, the thresholds are found be lower than the default threshold of 0.5. This choice of threshold is made to optimise the F_2 score, which places a higher emphasis on minimising false negatives compared to false positives. To reduce the number of false negatives, the optimal threshold will be lower than the default.

The performance of the models trained using the supervised contrastive learning regime relative to their baselines varies between the three architectures. In this regard, the ViT does not seem to be as suitable a candidate for contrastive learning as the ResNet model. This observation is illustrated

Table 4.1: Key evaluation results on the validation data partition, given batch sizes of 16 and 64. The threshold for classification optimises the F_2 score on the validation data.

Encoder		ViT		ResNet50		InceptionV3	
Encoder parameters		34,276,288		23,564,800		21,802,784	
Training method		Base	SupCon	Base	SupCon	Base	SupCon
Batch size: 16	Threshold	0.2455	0.2859	0.2131	0.2212	0.2212	0.2455
	AUC	0.8829	0.8759	0.8589	0.8889	0.8792	0.8740
	Accuracy	0.7109	0.7482	0.7728	0.7655	0.7568	0.6969
	Recall	0.9277	0.8735	0.8072	0.8916	0.8735	0.9398
	F_2	0.6210	0.6218	0.6052	0.6457	0.6288	0.6166
	Best score	2	2	1	3	3	1
	Mean score	0.7856	0.7798	0.7610	0.7979	0.7846	0.7818
Batch size: 64	Threshold	0.2939	0.1646	0.2455	0.2374	0.1889	0.3101
	AUC	0.8646	0.7604	0.8681	0.8750	0.8703	0.8741
	Accuracy	0.7295	0.5396	0.7688	0.6969	0.7022	0.7722
	Recall	0.8675	0.9458	0.7952	0.9277	0.9036	0.8373
	F_2	0.6040	0.5223	0.5951	0.6106	0.6024	0.6216
	Best score	3	1	1	3	1	3
	Mean score	0.7664	0.6920	0.7568	0.7776	0.7696	0.7763

in Table 4.1 where, under the ResNet architecture, supervised contrastive learning outperforms the baseline for three of the four metrics, and produces a higher mean metric score across both batch sizes. Conversely, under the ViT architecture, the mean metric score for models trained with supervised contrastive loss is lower than that of the baseline models. In contrast to the outcomes of the ViT and ResNet models, the Inception encoder exhibits mixed results. When employing a batch size of 16, the Inception baseline attains better performance, whereas supervised contrastive learning produces more favorable outcomes when the batch size is increased to 64. A further, detailed comparison of the encoders' performance is provided in Section 4.2.

Table 4.1 also illustrates that batch size may serve as an indicator of model performance within the given hyperparameter configuration. The table indicates that larger batch sizes result in slightly poorer mean metric scores across all architectures. This outcome is in contradiction with initial expectations, and this discrepancy is discussed further in Section 4.2.

Interpreting the 'best score' metric in Table 4.1, at a batch size of 16, the performance of the baseline models and the supervised contrastive learning models are tied. However, at an increased batch size of 64, the supervised contrastive learning regime yielded superior performance across two of the three encoder architectures. The significance of this result is discussed in Section 4.2.

Summing up the findings, three results stand out as significant.

1. The comparison between the supervised contrastive learning approach and the baseline ap-

proach yields mixed outcomes across the three encoder architectures. The supervised contrastive training does not consistently outperform, or underperform, the baseline model trained using binary cross-entropy.

2. On average, smaller batch sizes produce marginally better AUC, accuracy, recall and F_2 scores across both training methods.

3. Increasing the batch size appears to improve the performance of supervised contrastive learning relative to its baseline.

4.1.1 Test Results

The ResNet50 encoder architecture, trained using supervised contrastive learning and utilising a batch size of 16, emerges as the top performer in terms of validation AUC, F_2 metric, and mean score. As seen in Table 4.1, this configuration outperforms its baseline. Evaluating this model on the held-out test set yields:

- AUC: 0.8569
- Accuracy: 0.7055
- Recall: 0.8304
- F_2: 0.5717
- Mean score: 0.7411

As anticipated, the results on the test set are less favorable than the validation results as the classification threshold was determined using the validation data. However, the decline in performance is not considerable and the chosen model generalises sufficiently well to unseen data.

Samples from the test dataset are visualised in Figure 4.1 in the form of a confusion matrix. The arrangement of samples into true positives, true negatives, false positives, and false negatives provides a telling overview of the classifier's performance. The set of images falsely classified as non-melanomas, i.e. the false negative samples, exhibit less pronounced characteristics associated with melanoma relative to the rest of the samples in Figure 4.1. Conversely, the false positive segment showcases images that display indicators of melanoma according to the ABCD diagnostic criteria described previously. These hallmark traits of melanoma include asymmetry, irregular borders and varied colouring. The model demonstrates an impressive ability to identify melanomas that possess the appearance of ordinary lesions, as evident in the true positive block. Furthermore, it accurately categorises images as non-melanomas even when they exhibit features associated with melanomas. Ultimately, Figure 4.1 illustrates the model's proficiency and shows that the minimal

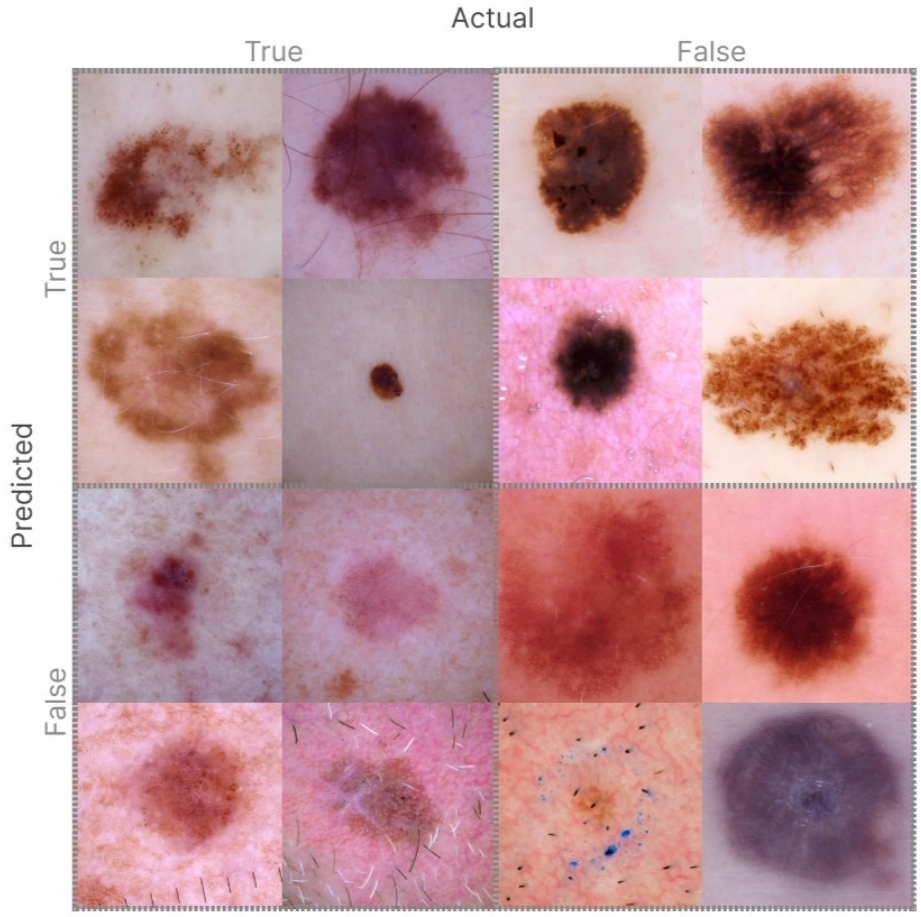

Figure 4.1: This figure presents samples of correctly and incorrectly classified melanomas from the test dataset. The snapshot provides insight into the model's effectiveness in distinguishing between melanomas and ordinary lesions.

errors it makes are understandable.

4.2 DISCUSSION

The three main findings are contrary to initial expectations. The first result indicates that the relative performance of the supervised contrastive training regime against its baseline is inconsistent across the three encoder architectures. The largest discrepancy in performance is between the ResNet architecture and the ViT. Under the ViT encoder, the SupCon experiments consistently produced lower mean scores than their baselines, whereas both SupCon experiments outperformed their baselines under the ResNet architecture.

Whether this finding is significant is uncertain. Given computational restrictions, none of the ar-

chitectures underwent hyperparameter tuning. Consequently, the default hyperparameters might have coincidentally suited the ResNet architecture better than the ViT architecture for the task of contrastive learning. Additionally, each encoder-batch size configuration was trained only once. The absence of repeated training and evaluation for each architecture introduces uncertainty regarding the superiority of any architecture over another. This uncertainty is accentuated by the inherent stochastic nature of neural network training and the selected batching methodology incorporating random, independent sampling. Existing literature also casts doubt on the validity of this finding. Amongst other articles, the papers mentioned in the literature review by Aladhadh et al. (2022), Xin et al. (2022), and Cirrincione et al. (2023) demonstrate significantly better melanoma detection performance using ViT architectures in comparison to earlier state-of-the-art CNN-based architectures. The existing body of literature, together with the lack of hyperparameter tuning and repeated experimentation calls the validity of this finding into question. Further experimentation and tuning should be employed to support the result, or refute it.

The aim of this research is to assess whether supervised contrastive learning outperforms traditional image classification models in the detection of melanoma. The inconsistent findings discussed above suggest that the performance of supervised contrastive learning compared to its baseline is sensitive to the encoder chosen as well as the parameter configuration of the model. Therefore, supervised contrastive learning will not consistently produce superior melanoma detection performance. A secondary aim of this study was to determine whether the enhanced long-range spatial awareness of ViT models yield superior melanoma detection results over CNN architectures. The study's findings suggest that ViT encoders do not necessarily outperform CNN architectures, particularly under the supervised contrastive training procedure. The ViT encoder with a batch size of 16 achieves the highest baseline performance; however, the performance gap is minimal. Additionally, it is important to note that the elevated performance should be evaluated within the context of the significantly larger number of encoder parameters in the ViT compared to ResNet50 and InceptionV3.

The second listed finding illustrated that smaller batch sizes produced marginally better performance across both training regimes. The mean score showed an average improvement of only 0.0253 when comparing small batch sizes to larger ones, suggesting that the performance gap could be negligible. Three hypotheses relating to this finding are suggested.

1. *The second finding may be refuted if extensive hyperparameter tuning is conducted and reveals that larger batch sizes produce better performance, as suggested in literature.* As mentioned in Section 2.1, larger batch sizes typically produce better results under a supervised contrastive training regime due to the nature of in-batch negative and positive sampling. In the same section, it was hypothesized that batch sizes as large as those in the original 'Supervised

Contrastive Learning' paper by Khosla et al. (2020) would be unnecessary given the binary nature of melanoma detection. Nonetheless, the assumption remained that larger batch sizes would lead to improved performance. The results in Table 4.1 are in contradiction to this prior-held assumption. One possible explanation for these results is that the default hyperparameter configuration was not altered to accommodate the larger batches. With a more fitting hyperparameter configuration, larger batches could potentially demonstrate improved performance, especially considering the slight performance difference observed in Table 4.1.

2. *The second hypothesis presented suggests that the effect of batch size cannot be accurately determined without considering batch sizes significantly larger than 64.* Another plausible hypothesis for this finding is that the performance gap is negligible given the minor difference in batch sizes employed in this study compared to the batch sizes presented by Khosla et al. (2020). In the 'Supervised Contrastive Learning' paper, results were reported using batch sizes up to 6,144 which increases to 12,288 after the double augmentation of each input. Due to a limited computational budget and memory capacity, the batch sizes considered in this study are dwarfed by those employed by Khosla et al. (2020). Therefore, a more in-depth investigation of the effect of batch size on classification performance is required to confirm or challenge the second finding of this study. This involves rerunning experiments with much larger batch sizes, coupled with comprehensive hyperparameter tuning tailored to each batch size configuration.

3. *The final hypothesis to the second finding is that smaller batches produce better validation performance on account of the implicit regularisation introduced by their reduced size.* The last hypothesis presented is that the results are not coincidental, and smaller batch sizes may yield optimal performance under a binary classification problem. As discussed, melanoma detection may not necessitate large batch sizes due to its binary nature. Under this assumption, smaller batch sizes may produce superior generalisation ability as suggested by the validation results in Table 4.1. Smaller batch sizes introduce noise into the training procedure which has a regularisation effect, preventing overfitting and producing superior results on a validation dataset (Kandel & Castelli, 2020).

The third, and final, main result states that increasing batch size improves the performance of the SupCon training regime in comparison to its baseline. The findings presented by Khosla et al. (2020) suggest that both cross-entropy loss and SupCon loss benefit equally from larger batch sizes. Illustrated in a figure within their study, it is evident that while the supervised contrastive loss outperforms cross-entropy across the full range of batch sizes employed, the improvements in top-1 accuracy increase with batch size at a similar rate for both methods. Notably, the smallest batch size considered by Khosla et al. (2020) is 512, or 1,024 after augmentation, whereas the largest batch

size considered in this research is 64. The findings in this research may indicate that, for batch sizes smaller than 512, supervised contrastive loss benefits from larger batch sizes to a greater extent than cross-entropy. However, this conclusion requires validation through comprehensive hyperparameter tuning and repeated experimentation.

Given the arguably minimal differences in results observed across the various encoder architectures, training methods, and batch sizes, the findings from this study require additional validation. Unfortunately, even with a relatively small batch size of 64, the training of a single encoder on the HAM10000 dataset using the two training methodologies consumes a significant amount of time and computational resources. Using a local computing setup, the highest attainable batch size was restricted to 16 due to memory allocation errors. These experiments were conducted on an Nvidia GeForce GTX 1650 GPU with four GB of dedicated memory. Consequently, further experimentation to explore larger batch sizes was conducted using a virtual machine equipped with 16 CPU cores (Intel(R) Xeon(R) Platinum 8168 CPU), 32 GB RAM, and 128 GB disk space. Despite the greatly increased memory capacity of the virtual machine, memory allocation errors arose for batch sizes larger than 64. Given the available resources, pursuing hyperparameter tuning, repeated experimentation and any other additional validation is unfeasible.

4.3 COMPARISON TO EXISTING LITERATURE

Comparing the findings of this research to existing literature presents a challenge due to the delayed publication of HAM10000 test labels. These labels were only made available in 2023 due to their utilisation in the 2018 ISIC melanoma detection challenge. While many modern papers discussed in Chapter 2 used HAM10000 for training, they reported results solely on the training dataset since the ground truth labels were not available at the time of publication. Training results do not constitute a fair basis of comparison as the models may be overfitted and consequently have poor generalisation ability.

The challenge hosts, however, shared the test results of the top five submissions in Codella et al. (2019). Among the top-performing models, the test AUC scores varied from 0.945 to 0.959, while the test accuracy ranged from 0.827 to 0.896. Considering the substantial advancements in the field of computer vision since 2018, including the introduction of the vision transformer, it is likely that modern algorithms may exceed the performance of the top challenge submissions. Notably, the top submissions significantly outperformed the results of this study and it is highly probable that modern algorithms would likewise surpass these results.

CHAPTER 5

LIMITATIONS AND SUGGESTED IMPROVEMENTS

In this chapter, the limitations of the study are examined, and suggested enhancements are presented to improve model performance, generalisability, and equitability. Previously mentioned improvements, such as the independently-sampled batching strategy outlined in Section 3.5 and the exploration of alternative pretext tasks in Section 3.1, are part of the suggested enhancements. This chapter elaborates on additional approaches to enhance the research project.

5.1 INCLUSION OF FITZPATRICK SCALE DURING INFERENCE

Section 3.1 describes the bias of dermatology datasets towards images of skin lesions on lighter skin tones. This section aims to describe an approach to account for the limited data available across diverse skin tones.

The Fitzpatrick scale is a classification system for human skin colours according to the level of pigmentation present in the skin. To improve this study, a suggested enhancement involves building a model that classifies skin lesion images according to the Fitzpatrick scale. A record of the representation of each pigmentation class in the training dataset used for melanoma detection can then be produced. During test time, the melanoma detection system can determine the Fitzpatrick classification of the lesion image and if the image corresponds to a Fitzpatrick class that is underrepresented in the training data, the model can provide a confidence score for the melanoma detection prediction. Additionally, for these test cases, the model could output a cautionary note about the potential risks associated with interpreting predictions for skin types that have not been frequently seen in the training data.

One such dataset that could be used to train a model to classify skin types is the Fitzpatrick 17k dataset made available by Groh et al. (2021). However, it is worth acknowledging that the Fitzpatrick 17k database suffers from class imbalance, with a higher abundance of images depicting light-skinned individuals compared to other skin types. An alternative approach to using neural networks for Fitzpatrick scale classification of lesion images is discussed in Groh et al. (2021). The authors suggest masking out regions of lesions or diseased skin in images based on the RGBA and YCbCr colour scales of the pixels. After masking, the colour of the remaining non-diseased skin can be aligned with the Fitzpatrick scale.

By comparing the performance of the neural network and masking approach, the best method can be determined and integrated into the melanoma detection system, as explained in the previous paragraph. By leveraging the Fitzpatrick scale, the melanoma detection system will aid patients and health practitioners in assessing the confidence level of the prediction.

5.2 HAIR REMOVAL DATA PRE-PROCESSING

To ensure the optimal generalisability of melanoma detection models and prevent irrelevant features (such as rulers in skin lesion images) from influencing the model's predictions, input images to a melanoma detection system should be standardised as much as possible. A number of researchers have developed machine learning models that learn to eliminate hair from lesion images which improves the standardisation of model input. In an article by Bardou et al. (2022), the authors propose a hair removal method for lesion images using a variational autoencoder. Unlike a large portion of the literature on this topic, their proposed algorithm does not require paired samples for its functionality. Incorporating this algorithm, or similar approaches, into the data processing step of this study has the potential to enhance the model's classification accuracy on both seen and unseen datasets. Including image hair removal is recommended for any future iterations of this research.

5.3 INCLUSION OF PATIENT METADATA

The HAM10000 dataset includes patient metadata including the patient's sex and age, the location of the lesion, and the method of diagnosis (i.e. histopathology, expert consensus or in-vivo confocal microscopy). The current model architecture classifies an image without considering the patient metadata, however, classification accuracy may be improved if these fields were considered. According to the study conducted by Bellenghi et al. (2020), it is evident that melanoma tends to have gender-specific behaviour with a higher occurrence in men. Additionally, melanoma is more frequently observed on the trunk of men and on the lower limbs of women. Furthermore, age is a well-established risk factor for melanoma, with the likelihood of melanoma occurrence increasing with patient age. In conjunction with numerous other articles, the study conducted by Bellenghi et al. (2020), suggests that age, sex and lesion localisation are relevant predictors for melanoma therefore including the HAM10000 tabular data could improve the detection ability of the classifier. The existing model architecture could easily be modified to include a MLP which encodes the tabular features. The results from both the tabular MLP encoder and the image encoder could be concatenated and then fed into either the projection head during the first phase of training, or the MLP during the second phase.

The present model configuration dynamically generates batches using an image generator. In the course of investigating techniques for producing batches containing corresponding image and tabular data, it became evident that most methods require loading the complete image and tabular dataset into memory prior to training, as opposed to generating them dynamically. Given the size and number of images, together with the existing computational restrictions of this research project,

loading the full image and tabular dataset into memory at once was not feasible. Nevertheless, with dedicated effort, the data generator can be adapted to dynamically produce aligned image and tabular data pairs. This modification has the potential to increase model performance but may lead to a slowdown in training times, as the process would involve the retrieval of the corresponding tabular data for each image.

5.4 EXTENSIVE HYPERPARAMETER TUNING

As mentioned numerous times in this study, this research project has a restricted computational budget. The computational limitations together with the scale of the dataset and encoders used in this study were such that hyperparameter tuning was not feasible. Hyperparameter tuning has the potential to significantly improve the model's performance and should unquestionably be utilised when sufficient computational resources are available.

CHAPTER 6

CONCLUSION

The purpose of this study is twofold. Firstly, this research aimed to compare the melanoma detection performance of models trained under the supervised contrastive learning regime to baseline models trained using binary cross-entropy loss. Secondly, this study aimed to determine whether vision transformers outperform traditional CNN architectures at the task of melanoma detection, as suggested in recent literature.

Prior to experimentation, the initial hypothesis was that employing the supervised contrastive loss and training approach proposed by Khosla et al. (2020) would improve classification performance relative to models trained end-to-end using cross-entropy loss. This hypothesis was grounded in the results of Khosla et al. (2020) and numerous other articles which showed that including a contrastive learning objective improved performance. Another hypothesis held prior to experimentation was that vision transformers would yield better melanoma detection results than traditional CNNs across both training regimes explored, as indicated by recent research findings. The findings of this study, however, did not align with either hypothesis. The experiments conducted using supervised contrastive learning yielded a mixture of results, with the training method not consistently demonstrating better or worse outcomes than the baseline approach. Additionally, the vision transformer did not demonstrate superior performance over popular CNN architectures considered.

The study encountered severe limitations due to insufficient computational resources and the heavy computational demands of the supervised contrastive learning training regime. These restrictions hinder the ability to adequately address the research questions posed. Consequently, further validation of the study's findings is required by means of repeated experimentation and extensive hyperparameter tuning should sufficient resources become available. Additional validation is critical considering the discrepancy between the study's results and the existing body of literature.

In the broader context, the results of this study highlight the complexities and challenges inherent in melanoma detection research. Notably, this study addresses the underrepresentation of darker skin tones within publicly available dermatology datasets. To promote equitable healthcare outcomes, a diverse skin tone dataset was utilised in the pre-training phase of the model. Furthermore, a proposition is put forth to assist medical practitioners in evaluating prediction confidence, based on the representation of skin tones in the training of melanoma classification models.

While the immediate outcomes of this research did not align with initial expectations, a host of recommended improvements to the design of the study have been proposed, coupled with suggestions to validate or refute the findings. The suggested enhancements are presented in such a way that may guide future research endeavors.

REFERENCES

Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S., and Islam, M. An effective skin cancer classification mechanism via medical vision transformer. *Sensors*, 22(11):4008, 2022.

Almaraz-Damian, J.-A., Ponomaryov, V., Sadovnychiy, S., and Castillejos-Fernandez, H. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy*, 22(4):484, 2020.

American Academy of Dermatology Association. What to look for: ABCDEs of Melanoma. `https://www.aad.org/public/diseases/skin-cancer/find/at-risk/abcdes`, 2023.

American Cancer Society. Cancer facts & Figures 2017. `https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html`, 2017.

American Cancer Society. Can Melanoma Skin Cancer Be Found Early? `https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/detection.html`, 2019.

Bahaa, M. KerasGen. `https://github.com/ma7555/kerasgen`, August 2022.

Bardou, D., Bouaziz, H., Lv, L., and Zhang, T. Hair removal in dermoscopy images using variational autoencoders. *Skin Research and Technology*, 28(3):445–454, 2022.

Bellenghi, M., Puglisi, R., Pontecorvi, G., De Feo, A., Carè, A., and Mattia, G. Sex and gender disparities in melanoma. *Cancers*, 12(7):1819, 2020.

Chen, K., Zhuang, D., and Chang, J. M. SuperCon: Supervised contrastive learning for imbalanced skin lesion classification. *arXiv preprint arXiv:2202.05685*, 2022.

Cirrincione, G., Cannata, S., Cicceri, G., Prinzi, F., Currieri, T., Lovino, M., Militello, C., Pasero, E., and Vitabile, S. Transformer-Based Approach to Melanoma Detection. *Sensors*, 23(12):5677, 2023.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Cueva, W. F., Muñoz, F., Vásquez, G., and Delgado, G. Detection of skin cancer 'Melanoma' through computer vision. In *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pp. 1–4. IEEE, 2017.

Daneshjou, R., Vodrahalli, K., Novoa, R., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S., Bailey, E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J., Okata-Karigane, U., Zou, J., and Chiou, A. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances DOI: 10.1126/sciadv.abq6147*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828, 2021.

Gupta, K., Ajanthan, T., Hengel, A. v. d., and Gould, S. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022.

Hsu, B. W.-Y. and Tseng, V. S. Hierarchy-aware contrastive learning with late fusion for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 216:106666, 2022.

Ichim, L. and Popescu, D. Melanoma detection using an objective system based on multiple connected neural networks. *IEEE Access*, 8:179189–179202, 2020.

Jain, S., Pise, N., et al. Computer aided melanoma skin cancer detection using image processing. *Procedia Computer Science*, 48:735–740, 2015.

Kaggle. SIIM-ISIC Melanoma Classification. `https://www.kaggle.com/competitions/siim-isic-melanoma-classification/overview`, 2020.

Kandel, I. and Castelli, M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT express*, 6(4):312–315, 2020.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Kundu, R. The Beginner's Guide to Contrastive Learning. https://www.v7labs.com/blog/contrastive-learning-guide, 2023.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., and Wang, C. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics*, 71:19–29, 2019.

Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Dorffner, G., and Ellinger, I. Investigating and exploiting image resolution for transfer learning-based skin lesion classification. In *2020 25th international conference on pattern recognition (ICPR)*, pp. 4047–4053. IEEE, 2021.

Masood, A. and Ali Al-Jumaily, A. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International Journal of Biomedical Imaging*, 2013, 2013.

National Cancer Institute. What Does Melanoma Look Like? https://www.cancer.gov/types/skin/melanoma-photos, 2011.

Phillips, M., Greenhalgh, J., Marsden, H., and Palamaras, I. Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatology Practical & Conceptual*, 10(1), 2020.

Popescu, D., El-Khatib, M., El-Khatib, H., and Ichim, L. New trends in melanoma detection using neural networks: a systematic review. *Sensors*, 22(2):496, 2022.

Shehzad, K., Zhenhua, T., Shoukat, S., Saeed, A., Ahmad, I., Sarwar Bhatti, S., and Chelloug, S. A. A Deep-Ensemble-Learning-Based Approach for Skin Cancer Diagnosis. *Electronics*, 12(6): 1342, 2023.

Thanh, D. N., Prasath, V. S., Hieu, L. M., and Hien, N. N. Melanoma skin cancer detection method based on adaptive principal curvature, colour normalisation and feature extraction with the abcd rule. *Journal of Digital Imaging*, 33:574–585, 2020.

Tschandl, P. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. https://doi.org/10.7910/DVN/DBW86T, 2018. Harvard Dataverse, Version V4, DOI: 10.7910/DVN/DBW86T.

Tschandl, P., Rosendahl, C., and Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., Zhou, Q., Wang, S., Li, L., Yang, F., et al. An improved transformer network for skin cancer classification. *Computers in Biology and Medicine*, 149:105939, 2022.