INTEGRATING EXPLAINABLE AI FOR TRANSPARENT SKIN LESION CLASSIFICATION

ARISSA NOORDINA BAHARI

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2024

ABSTRACT

The accurate classification of skin lesions in medical imaging is a critical task, necessitating advanced machine learning solutions. While existing models often utilize deep neural networks to achieve high accuracy, they frequently lack transparency and interpretability, limiting their clinical utility. This research aims to bridge this gap by integrating Explainable Artificial Intelligence (XAI) techniques into skin lesion classification systems. Utilizing an MLP-MobileNetV3 architecture for robust image feature extraction, combined with colour histogram analysis and patient metadata, can enhance model performance. Additionally, interpretability is addressed through the application of saliency maps, providing visual insights into the model's decision-making process. This study employs the ISIC 2019 Challenge dataset, consisting of 25,331 images categorized into eight classes, and incorporates detailed preprocessing steps including image augmentation, metadata encoding and colour histogram creation to further improve classification accuracy. Despite achieving a best accuracy of 40% and a mean sensitivity of 54%, the MLP-MobileNetV3 model falls short of the baseline performance of 63% accuracy and 73% sensitivity. The results indicate a potential reduction in false positives and underscore the importance of model interpretability, particularly in clinical settings. Future work should address model limitations, extend training epochs, and enhance data handling practices to advance skin lesion classification and its practical applications in healthcare.

Keywords: Explainable skin lesion classification, deep learning, MobileNetV3, saliency maps

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my supervisor, Dr Erma Rahayu, without whose insight this research report would not have been possible. The continuous collaboration, understanding, patience and kindness shown by her empowered me to further my report in the field of explainable skin lesion detection. She has facilitated my interest and growth in the field of computer vision, allowing me to reach greater heights.

I would also like to thank my classmates and peers, without whom my Masters journey would not have been the same. With each other, we found great moral support, collaborative spirit and camaraderie in our shared frustrations and late-night meals.

Lastly, I want to thank my family, especially my mother and sister. They were my motivation in times of extreme stress and my biggest supporters, fully believing in me until the end. I thank my father too, who although isn't here anymore, who I know is looking over me and guiding me in spirit. And finally, thank you to my dear boyfriend Malcolm, who has been my rock and my cheerleader, helping me both in coding and outside of it.

I cannot thank you all enough.

Table of Contents

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ACC	:	Accuracy
AK	:	Actinic kerastosis
BCC	:	Basal cell carcinoma
BKL	:	Benign keratosis lesion
CNN	:	Convolutional Neural Network
DF	:	Dermatofibroma
F1	:	F1-score
MEL	:	Melanoma
MLP	:	Mutilayer perceptron
NV	:	Melanocytic nevus
PRE	:	Precision
REC	:	Recall
SCC	:	Squamous cell carcinoma
SEN	:	Sensitivity
SPE	:	Specificity
SVM	:	Support Vector Machines
VASC	:	Vascular lesion
XAI	:	Explainable AI

1. Chapter 1: Introduction

1.1 Research Background

Skin lesions, ranging from benign growths to malignant cancers, represent a significant challenge in medical diagnostics, exacerbated by their diverse appearances and potential health implications. Melanoma and non-melanoma skin cancers are among the most prevalent and concerning types, necessitating accurate and timely diagnosis for effective treatment. Skin cancers remain the most common group of diagnosed cancers across the globe, with an estimation of 330,000 new cases of melanoma alone diagnosed globally (WHO, 2022). Early detection is key; new melanomas are shallower and thinner than those that have metastasized, decreasing treatment difficulty.

Skin lesions can be categorized into several classes, including Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic Keratosis (AK), Benign keratosis lesion (BKL), Dermatofibroma (DF), Vascular lesion (VASC), and Squamous cell carcinoma (SCC). Generally, skin cancers develop from ultraviolet (UV) ray exposure from the sun, as well as artificial lighting from sunlamps and tanning beds. While the majority of skin cancers frequently occur and are easily treated, melanoma (MEL) constitutes over 70% of skin cancer fatalities despite representing around 5% of all skin cancers. With its propensity to grow and spread further than other types of skin cancers, its early detection is paramount to saving lives.

Challenges in accurately classifying skin lesions stem from their variability in appearance due to factors such as skin type, lesion morphology, and imaging conditions. Even trained dermatologists find difficulty in accurately diagnosing skin lesions, with only extensive training and experience leading to better diagnoses. Achieving robust classification with artificial intelligence thus requires advanced feature extraction techniques and machine learning algorithms capable of discerning subtle differences crucial for diagnosis. Explainable AI (XAI) methods play a pivotal role in enhancing transparency and interpretability in dermatological diagnostics due to the black-box nature of the diagnostics models. By employing techniques such as feature visualization, XAI elucidates the reasoning behind AI-driven diagnostic decisions. This transparency not only builds trust among healthcare providers but also facilitates the integration of AI systems into clinical practice by empowering clinicians to understand and validate AI recommendations. Collaborative efforts, such as those supported by initiatives like ISIC, are crucial in advancing the field towards more reliable and accessible diagnostic tools for skin diseases.

This research report delves into the intersection of XAI and skin lesion classification with an aim to bridge the gap between highly advanced deep learning models and user understanding. By making clear the decision making processes of the deep learning models with XAI, this report aspires to develop an accurate and explainable skin lesion classification system.

1.2 Problem Statement

The challenge of skin lesion classification lies in the need for accurate and timely diagnosis, particularly in scenarios where traditional methods may be limited by subjectivity, resource constraints, or lack of specialized expertise. While advancements in machine learning have made models more sophisticated and facilitated automated skin disease detection from images, the black-box nature of these models hinders their integration into clinical practice. Models should be integrated with Explainable Artificial Intelligence (XAI) techniques to enhance transparency, interpretability and model accountability.

The current landscape in skin lesion detection relies heavily on deep learning models, particularly pretrained Convolutional Neural Networks (CNNs) to classify images. While these models demonstrate high accuracy, the lack of transparency in their decision-making processes due to the complex and black-box nature hinders a user's ability to trust the model's classification rationales. The integration of XAI elements into the skin lesion classification systems can prove insightful into the patterns the influence a model's decision, which in turn will improve user confidence, and reduce potential bias and harms from the model.

Thus, there is a critical demand for an explainable skin lesion detection system that not only delivers precise diagnoses but also offers transparent explanations for its decisions. Such a system would enhance trust among healthcare providers and patients, streamline diagnostic processes, and pave the way for personalized treatment strategies, ultimately improving patient outcomes in dermatological care.

This research report aims to explore and develop innovative methods of integrating XAI modules within skin lesion classification models. The goal is to create a robust, highly interpretable and trustworthy system that accurately classifies skin lesion and provides transparent and direct explanations for its classification. Using colour histograms, metadata integration, MobileNetV3 deep learning model and saliency map visualisations, this architecture may be a step closer to solving the issue of skin lesion classification. As such, we will use the above architecture in the paper.

1.3 Research Questions

Three research questions are highlighted to understand the XAI methods used explainable skin lesion classification:

1. What are the machine learning and deep learning methods used for skin lesion classification?
2. Which features are most relevant for effective skin lesion classification?
3. How can explainable AI models be effectively integrated for explaining the decisions of skin lesion classification models?

1.4 Research Objectives

In line with the previous section on research questions, the following research objectives are established:

1. To review the machine learning and deep learning methods used for skin lesion classification.
2. To identify the most relevant feature extraction in skin lesion classification
3. To integrate an effective explainable AI model most suitable for explaining the decisions of skin lesion classification models.

1.5 Research Significance

Skin diseases and lesions pose a significant threat to public health due to their potential to develop into serious conditions, such as melanoma. These conditions are particularly concerning because they can lead to severe health complications, including disfigurement, systemic infections, and even death if not diagnosed and treated promptly. For example, melanoma, a type of skin cancer, is highly aggressive and can spread rapidly to other parts of the body, making early detection and treatment crucial. While the detection of skin diseases are paramount, explainable skin disease methods are equally important in fostering trust in AI systems. Without concrete explanations on my models classify diseases into specific categories, belief in AI medical systems will erode, along with its decrease in usage. As such, this research report involves collecting and analysing skin lesion data, implementing a deep learning model to classify the images, and examining the best explainability model to explain the decision of an AI model. This report aims to build an accurate, trustworthy and interpretable predictive model that can help users identify different skin lesions within an acceptable timeframe. Better understanding of the methods, algorithms and explainability modules of skin lesion classification will therefore be important in understanding the needs of future explainable skin lesion detection systems.

2 Chapter 2: Literature Review

Since the International Skin Imaging Challenge (ISIC) began in 2016, research on skin imaging and skin lesion classification has steadily increased with time. The integration and accessibility of AI, along with the start of the ISIC competitions have made research within this topic more accessible for many around the world. With the growing interest in explainable AI (XAI), many papers have chosen to incorporate explainability models into skin lesion classification.

In this chapter, we aim to explore the literature within the field of explainable skin lesion classification, provide a comprehensive critical literature review discussion table to visualise the exploration, summarise the related work and provide a background to the proposed method. We discuss three approaches to explore the literature: general papers on skin lesion classification, explainable skin lesion classification papers, and papers related to the ISIC 2019 dataset we will be using for this project. All papers within the review were published between 2019 to 2023.

2.1 Papers on skin lesion classification

In this section, we explore 4 papers that take different approaches on classifying skin lesions for 4 different datasets. These aim to give a brief overview on the literature.

In their work, Allugunti (2022) experimented on the DermNet dataset to classify three melanoma types: lesion maligna, superficial spreading and nodular melanoma. Amongst the decision tree, random forest, gradient boosted trees and CNN classifier models, the CNN model achieved the best results with precision, recall, F1-score and accuracy scores of 91.07%, 87.68%, 89.32% and 88.83% respectively. While the model attained high metric scores and looks promising for multiclass classification, the author did not include details on the CNN model architecture, feature representations and experimental setup. Future work should include these details for better reproducibility.

Ahmad et al. (2020) fine-tuned ResNet152 and InceptionResNet-V2 models with a triplet loss function to classify skin lesions from the AI Skin dataset. Input images are mapped to a 128-D Euclidean space, and L-2 distances between embeddings are calculated to compare images. By utilising the triplet loss function, embeddings of images from the same class are closer together, and distinct from images of other classes. Classifying four categories, the InceptionResNet-V2 + Triplet model achieved the better scores between the two, with an accuracy of 87.42%, recall of 97.04% and specificity of 96.48%. While the model outperformed the state-of-the-art works for skin disease classification, it could be improved with a better dataset curated by a dermatologist. The small dataset was only made up of 800 total images consisting of vague categories such as acne, blackheads, dark circles and spots.

Focusing on the effects of texture and colour on classification accuracy scores, Swamy and Divya (2021) used classical machine learning algorithms of decision trees and SVM models on the DermNet and DermQuest datasets. Overall, the decision tree model achieved accuracies of 66% and 75% respectively when tested of colour and texture features, while SVM accuracies were 75% and 83% respectively for colour and texture features. Despite the simple algorithms used for the experiment, texture feature extraction proved to improve and achieve better results than colour features. However, the accuracies achieved are much lower than existing state-of-the-art, with better feature extraction methods, such as combining texture and colour extractions, necessary to improve results.

Operating on the Xiangya-Derm dataset containing 6 common skin diseases, Wu et al. (2019) tested 5 different CNN models, including ResNet50, Inception V3, DenseNet-121, Xception, Inception-ResNet V2 to conclude which was the best in classifying skin disease. Experiments were performed without mention of fine-tuning. Their experiments highlight that models previously trained on differing body part images performed better than those that weren't, despite the experiment dataset containing only facial skin images. Of the 5 models, InceptionResNet-V2 achieved the best precision and recall scores of 70.8% and 77.0%. Overall, the paper showed that pretrained CNN architectures achieved satisfactory results with minimal finetuning. However, the dataset's quality must be improved by adding more images within the database, and increasing model finetuning to achieve better results, as the best results score were still quite low for the AK class of images.

2.2 Explainable skin lesion classification papers

In this section, we discuss explainable skin lesion classification papers found within the literature space. This section includes 7 papers that incorporate a range of explainability modules within their architecture.

Ahmad et al. (2023) focused their experiments on feature selection, model fusion and visualisation on the ISIC 2018 and HAM10000 datasets. Features were selected using Butterfly Optimisation Algorithm (IBOA) with the features then fused, while they chose the models Xception, ShuffleNet and a fusion model made from the two. To visualise the prediction, GradCAM was utilised. From their investigations, they achieved accuracy, recall, precision and F1-scores of 99.3%, 99.38%, 99.4% and 99.38%. While metrics were highly scored for both the HAM10000 and ISIC 2018 datasets, the models could be optimised better, such as by using Bayesian optimisation.

In their work, Ballari et al. (2022) experimented on a Kaggle skin disease dataset with ResNet-18 to classify skin disease images, and used GRAD-CAM as the explainability module. Results include 96% accuracy in classifying images, and a convolutional feature map displayed as the image output

highlighting the areas relevant to classification. While the results are visually interpretable, this paper is riddled with deficiencies, such as vaguely communicated results by the author and an unspecified dataset with unknown sizes. Future work must include these details for clarity and reproducibility.

Two works focus their attention to the ISIC 2017 and ISIC 2018 dataset (Barata et al., 2021; Ding et al., 2023). In Barata et al. (2021), their model system mimicked the hierarchical decisions made by dermatologists during skin lesion classification, such as the lesion origin, malignancy degree and the differential diagnosis, with each decision based on the previous one. The architecture include an image encoder block, either VGG-16, ResNet-50 or DenseNet-161, extracting discriminative features, a decoder block for hierarchical classification and a trainable attention module for explainability. The VGG-16 model version proved the best with sensitivity, specificity and AUC scores of 86.7%, 87.1% and 92.4% respectively. Their hierarchical taxonomy element bred competitive results as the model correctly identified relevant lesion region and colour normalisation improved accuracy. However, melanoma class was not easily identified with the model, and it was not robust withstanding varying image transformations such as rotation and scaling.

In contrast, Ding et al. (2023) made use of HI-MViT , a model based on the MobileViT block to improve classification accuracy and increase interpretability due to the transformer's self-attention mechanism. Their dataset also included the HAM10000 dataset, in addition to ISIC 2017 and ISIC 2018. To visualise the results, GradCAM and AblationCAM were used. Best results were attained on the ISIC 2018 dataset, with precision 93.1%, recall 93.2%, F1-score 93.1% and accuracy of 93.2%. Key strengths of this model are its high results, even in comparison to the state-of-the-art, and its generalisability on other datasets. However, the data trained and tested on lacked a diverse range of skin tones and lack other patient medical indicators, which can lead to bias and misclassification.

Employing a different explainability method altogether, Tschandl et al. (2019) used Content Based Image Retrieval (CBIR) to retrieve dermatoscopic images that were the most visually similar to the classification of an image, which was done using ResNet-50. Using data from EDRA, ISIC 2017 and their own private data, best classification results include accuracy of 76.2%, an AUC score of 85.0%, specificity scores of 92.2% and sensitivity scores of 72.7%. CBIR as an explainability method is intuitive as similar content as the classified image is retrieved, allowing observers to immediately spot the similarities. However, the paper acknowledges its limitations due to the dataset, and lack of finetuning of the ResNet-50 model as their focus was on CBIR mostly.

In Young et al. (2019) paper, they used the Inception CNN to classify skin lesion images, along with KernelSHAP and GradCAM for explainability. Their dataset consisted of 6017 images with significant class imbalances. Training 30 models with a range of learning rate, dropout and epoch

number, their mean AUC achieved was 85% with a mean recall of 87%. The use of both GradCAM and KernelSHAP allowed the authors sanity checks on two fronts, with the two different modes of explainability. However, the authors admit that they were limited by the small dataset size, which led to spurious correlations by the model. Furthermore, only testing the Inception architecture limited the accuracy ceiling.

Using images from an unspecified ISIC dataset, Zia Ur Rehman et al. (2022) trialled their experiment with MobileNetV2, chosen for its feature extraction capabilities, and DenseNet201 architectures. They also used GradCAM to visualise the results. Classifying benign and malignant melanoma, their best results came from the DenseNet201 model, with an accuracy of 95.50%, precision of 97.02%, F1-score of 95.46%, sensitivity of 95.96% and specificity of 97.06%, highlighting the high metrics achieved by the model. However, their work could be further improved by including better optimisation methods.

2.3 ISIC 2019 Related papers

In this section we focus on ISIC 2019 related papers. International Skin Imaging Competition (ISIC) 2019 is an annual international competition that focuses on skin imaging techniques and classification. In this 2019 iteration, contestants were tasked to classify skin lesions into 8 categories, with or without the inclusion of additional patient metadata. This section of the literature review focuses on 11 texts that utilise this dataset.

In their work, El-Khatib et al. (2020) focused solely on image data from the ISIC 2019 and PH2 datasets to classify skin lesions into two classes, melanoma and common nevus. They used multiple CNNs like GoogLeNet, ResNet-101, NasNet-Large and a decision fusion of all 3 models, and extracted features with histogram of gradients (HOG). The decision fusion model on the ISIC 2019 dataset brought about the best results, with accuracy 93.00%, specificity of 93.33% and sensitivity of 92.50%. While the metrics scored were high and the fusion model performed better than individual CNNs, the small size of the dataset (only 300 images) can lead the model to overfit on the data. Future work should include a larger dataset size.

With their submission which won the ISIC 2019 challenge, Gessert et al. (2020) made use of both the image dataset and the metadata dataset available to classify all 8 skin lesions classes. EfficientNets B0 to B6 were their CNNs of choice, while the metadata was input to a two layer neural network with ReLU activation and dropout. The information from both architectures were concatenated and ensembled before passing through a classification layer. Including the metadata, the

best results were AUC, AUC-S, sensitivity and specificity of 98.00%, 96.50%, 55,60% and 99,30% respectively. The authors agree that including metadata generally increase specificity and AUC scores across all 8 classes, but relent that the model does not work reliably on out of distribution images, and that including metadata decreased model sensitivity.

Gong et al. (2020) included StyleGAN generated images of multiple classes along with the ISIC 2019 dataset for their experiment. Experimenting with 43 different CNNs including AlexNet, BN-Inception, Inception, NasNet and ResNet, they ultimately went with a decision fusion model that combines the decision of multiple classifiers. Their best model, DecisionFusion3, achieved an accuracy of 99.50%, AUC of 98.90%, precision of 98.4%, sensitivity of 98.3% and specificity of 99.6%. Their work demonstrated that GANs alleviated the issues of smaller datasets and imbalanced classes as new data is generated, and the fusion model achieved high accuracy. However, with 43 different CNN model, selecting the best combination of fusion CNNs proved difficult.

Proposing a novel entropy-based weighting and first-order cumulative segmentation method to segment skin lesions from an image, followed by Wide-ShuffleNet to classify images from the ISIC 2019 and HAM10000 datasets, Hoang et al. (2022)'s experiment resulted in an accuracy of 82.56%, with sensitivity and specificity of 82.56% and 97.51% respectively. Their proposed framework is much lighter than different models tested such as VGG-19, which had 79 times higher parameters. Furthermore, this segmented model had higher accuracy than non-segmented models, and does not require a ground truth image. However, it performed slightly worse than the EW-FCM + EfficientNetB0 model.

Incorporating 3 ISIC datasets from 2017 to 2019, Iqbal et al. (2021) proposed their own model, Classification of Skin Lesion Network (CSLNet) comprised of four key kernel units with differing numbers of convolutional layers, between 9 – 27 layers. The units were key in detecting complex patterns within the images such as starburst and cobblestone patterns, as well as complex lesion features like cysts and pigment blotches. Their model resulted in best results for the ISIC 2019 dataset, with accuracy, precision, sensitivity, specificity, F1-score and AUROC scores of 89.58%, 90.66%, 89.58%, 97.57%, 89.75% and 99.1% respectively. Their results proved that the model consistently outperformed other state-of-the-art models for each dataset, with the different kernel units able to recognize symmetry, colour and complex patterns. However, the lack of further information inclusion in their dataset, such as metadata like age, race, and gender information could hinder their results.

Meanwhile, Kassem et al. (2020) proposed two architectures for the ISIC 2019 dataset to classify 8 different classes of skin images: a modified GoogLeNet with the last three layers removed, and GoogLeNet with bootstrapped SVM, where the last two layers of the CNN are replaced by SVM.

Their GoogLeNet only model performed better between the two, achieving an accuracy score of 94.92%, sensitivity scores of 79.8%, specificity scores of 97% and precision scores of 80.36%. Their model outperformed the winning model of the ISIC 2019 competition (Gessert et al., 2020), but the multiclass SVM iteration gave lower performance measures on the test set.

Two works include explainability into their models. Incorporating XAI into their unspecified ResNet classification model, Metta et al. (2021) utilised ABELE, an image classifier specific local model agnostic explainer. ABELE's explanations consist of exemplar and counter-exemplar images and a saliency map. The paper's focus is on explainability rather than results, and only published a balanced, multiclass accuracy of 0.838. When conducting their survey with real user experts on the ABELE web interface module, the survey supports that without consistent validation by real world experts, explanations are not useful. As such, future work into XAI for not just skin lesions classification need real world validation as well. Meanwhile, Nigar et al. (2022) used ResNet as their classifier as well with ResNet-18. Their explainablity method however is LIME, a model agnostic explainability method which provides local explanation for individual predictions by approximating decision boundaries within an image. The ResNet-18 classifier results were accuracy, precision, recall and F1-scores of 94.47%, 93.57%, 94.01% and 94.45% respectively. Their model proved its capabilities in generalizing well with both the model architecture and LIME's model agnostic nature, but with limitations including the lack of opposing examples, such as healthy skin, fingers, nose and eyes for contrast.

Olayah et al. (2023) employed geometric active contour (GAC) segmentation to isolate diseased skin from healthy skin, and then experimented with two different architectures: CNN-ANN and CNN-RF, with the CNNs including AlexNet, GoogLeNet and VGG-16. They also included hybrid CNN models in their experimentation. The best model, the hybrid AlexNet-GoogLeNet-VGG-16-ANN scored highly across all metrics with an accuracy of 96.10%, sensitivity of 88.90%, specificity of 99.44%, precision of 88.69% and AUC of 94.41%. The models used are optimized, with segmented and hybrid models achieving high accuracies. However, the dataset's imbalance produced limitations for the work and a fusion of handcrafted and CNN detected features could improve future work. Tô et al. (2019) submission to the ISIC 2019 challenge include the use of the metadata as well as the images and employed a segmentation-CNN methodology as well. They first segmented their dataset with the U-Net algorithm and classified the dataset with the EfficientNet-B4 CNN. They did not post their results of their experiments.

Finally, Villa-Pulgarin et al. (2022) classified images from both the ISIC 2019 and HAM10000 datasets into 7 classes. Experimenting with DenseNet-201, Inception-V3 and Inception-Resnet-V2 CNN

models, the former model proved to achieve highest metric scores of accuracy, precision, recall and F1-scores of 93% for all metrics. While the DenseNet-201 model is optimized compared to others in the literature, the minimal data preprocessing on the dataset can negatively affect the results. Future work can consider segmentation and contrast enhancement to improve accuracy in detection.

2.4 Literature Review Discussion

To start, with the exception of Allugunti (2022); Swamy and Divya (2021)'s work which included traditional machine learning models such as decision trees, SVM, random forest and gradient boosted trees, all other studies incorporated deep learning architectures for their models; more specifically, pretrained CNN models. For the non-ISIC 2019 related dataset papers, Ahmad et al. (2023)'s architecture of a fusion model of Xception and ShuffleNet achieved the highest model accuracy of 99.3% on the HAM10000 dataset. Across the papers based on the ISIC 2019 dataset, Gong et al. (2020) results were the best with their DecisionFusion3 model's accuracy of 99.50%. However, the winner of the ISIC 2019 challenge, Gessert et al. (2020) also scored highly across the metrics of AUC, AUC-S, sensitivity and specificity, with a maximum AUC score of 98.00%

Across the 22 papers analysed, only 4 papers explicitly noted methods of feature extraction. Colour extraction was featured in all 4 papers except Ahmad et al. (2020), which featured embeddings that were automatically calculated by the CNN model. Meanwhile, Swamy and Divya (2021) also included texture feature extraction in their model. The rest of the reviewed literature did not include feature representations; if they did, it was only stated that the CNNs would automatically extract the features during training.

Amongst the various CNNs used for classification, the top five most popular models are ResNet with 7 appearances, DenseNet with 4 appearances and Inception, Inception-ResNet and GoogLeNet with 3 mentions each. ResNet's popularity within the literature could be a result of the different architectures within the model, such as ResNet-18, ResNet-50 and ResNet-101. Fusion models were also popularly used by authors; indeed, the best performing overall model achieved this feat by utilising a fusion model of multiple CNNs. Furthermore, 3 papers made use of combining segmentation to isolate lesions in an image, and CNNs to classify it (Hoang et al., 2022; Olayah et al., 2023; Tô et al., 2019). The authors of said papers highlighted the optimised performance of the models, as the segmentation yielded better results for experiments than non-segmented models.

With regards to explainability, 7 different explainability models were used: GradCAM, attention modules, AblationCAM, kernelSHAP, CBIR, ABELE and LIME. From these 7 models, GradCAM was the most popular, appearing in 3 papers; other models only appeared in 1 paper each. GradCAMs are a

visualisation technique used on deep learning models, specifically for computer vision. They produce a heatmap that highlights the important regions of an image using the target gradients (such as class) as the final convolutional layer. As such, users can interpret the GradCAM visualisation as highlighting the final decisions on why a model classified an image as such. In the case of skin lesion classification, this is a direct way that highlights which region of an image relate to the particular lesion, and increases model explainability and interpretability. Ahmad et al. (2023) model resulted in the best classification accuracy for papers that included GradCAM.

Despite the strengths in the previous work in skin lesion classification models, limitations arise within the work. A common limitation mentioned surrounds the datasets used for experimentation. Multiple authors discuss that the datasets used are highly imbalanced, such as the ISIC 2019 dataset, as the melanoma nevus class makes up 73% of the 8 class dataset, which can lead to overfitting and bias from the models used. Furthermore, authors also note the lack of diversity of images within the dataset, such as the exclusion of various skin tones with skin lesion and patient metadata. Indeed, the lack of inclusion of additional metadata information, such as patient age, gender and race, used within experiments is notable, as only 1 paper (Gessert et al., 2020) included metadata in its classification model. This can limit classification capabilities. Datasets can also be improved with the help of dermatologists when curating them. In the papers related to the ISIC 2019 dataset, the low number of explainable models incorporated in papers also present an area of future improvement and research.

In conclusion, more diverse datasets are necessary to run the entire gamut of skin lesion classification tasks, as well as better datasets including patient metadata information. Future work that utilise robust, diverse datasets, as well as ensemble architectures with explainable models can improve model classification accuracy, interpretability and explainability.

In this work, we implement colour histogram feature representations, include patient metadata, use a MobileNetV3 architecture and provide saliency map visualisations to classify skin lesions.

2.5 Theoretical background

2.5.1 MobileNetV3

MobileNetV3 (Howard et al., 2019) is a neural network architecture specifically designed to operate efficiently on mobile and edge devices, where computational resources and power consumption are limited. It builds upon the foundations laid by its predecessors, MobileNet and MobileNetV2, focusing on optimizing both speed and accuracy for tasks like image classification.

MobileNetV3 utilizes a combination of advanced techniques to achieve this balance. At its core, the architecture employs depthwise separable convolutions, a key feature of the MobileNet family, which significantly reduces the number of parameters and computations compared to standard convolutions. This allows the network to process images with less computational overhead while still capturing essential features.

One of the notable innovations in MobileNetV3 is the use of neural architecture search (NAS) to automate the design of the network. NAS explores various possible configurations and optimizes the architecture for the best performance given the constraints of mobile devices. As a result, MobileNetV3 introduces lightweight attention mechanisms, such as the squeeze-and-excitation (SE) module, which helps the network focus on important features in the image without adding significant computational burden.

Furthermore, MobileNetV3 incorporates efficient building blocks like the hard swish activation function and the use of fewer layers in certain parts of the network, further reducing the model's complexity while maintaining accuracy. These optimizations make MobileNetV3 a powerful choice for image recognition tasks on mobile devices, providing a good trade-off between efficiency and performance.

2.5.2 Saliency maps

Saliency maps are a visualization technique used to understand which parts of an input image a neural network considers most important for making a classification decision. Introduced by Simonyan et al. (2013), saliency maps work by computing the gradient of the output score (e.g., the score for a specific class) with respect to the input image. This gradient information reveals how much a small change in each pixel would affect the classification score, effectively highlighting the regions of the image that are most influential in the model's decision.

The resulting saliency map is typically a grayscale image where brighter pixels indicate areas that have a stronger influence on the network's prediction. Unlike more complex techniques like Grad-CAM, which focus on specific layers within the network, saliency maps provide a more direct and straightforward visualization by operating at the input level. This simplicity makes them a valuable tool for quickly assessing which features of an image contribute most to a model's decision.

Saliency maps are particularly useful for interpreting and diagnosing the behaviour of deep learning models, especially in cases where understanding the model's focus can lead to better insights

into its performance and potential biases. By visualizing these maps, researchers can gain a better understanding of how the model processes information and whether it aligns with human intuition.

2.6 Critical Analysis Summary Table

In this section, we discuss the critical analysis of 22 papers that relate to skin lesion classification. These critical analysis tables function to identify reoccurring datasets used, various methods of feature extraction, algorithms included, explainability models, results, strength and weaknesses of the 22 works from the past 5 years. This table is key to understanding the current skin lesion classification landscape, and for future work to improve current model performance. Table 1 summarises 4 studies related to skin lesion classification, while Table 2 summarises information on explainable skin lesion works and Table 3 summarises literature related to the ISIC 2019 dataset.

Table 1: Summary of related studies to skin lesion classification

Article Information (APA)	Dataset	Feature representation	Algorithm	Results	Strength	Weakness
Allugunti, V. R. (2022). A machine learning model for skin disease classification using convolution neural network. *International Journal of Computing, Programming and Database Management*, *3*(1), 141-147.	Dermnet	-	Decision trees, Random Forest, Gradient Boosted Trees, CNN	Best results: CNN PRE: 91.07% REC: 87.86% F1: 89.32% ACC: 88.83%	High precision, recall, accuracy, and F1-scores Model looks promising for multiclass classification	Unclear feature representation
Ahmad, B., Usama, M., Huang, C. M., Hwang, K., Hossain, M. S., & Muhammad, G. (2020). Discriminative feature learning for skin disease classification using deep	AI-skin	Embeddings calculated by CNN and triplet loss function	ResNet152 + Triplet, Inception ResNet-V2 + Triplet	Best results: Inception ResNet-V2 + Triplet ACC: 87.42% REC: 97.04% SPE: 96.48%	Model outperforms SOTA works for skin disease classification	Model can be improved with a better dataset curated with dermatologist

convolutional neural network. *IEEE Access*, *8*, 39025-39033.						
Swamy, K. V., & Divya, B. (2021, December). Skin disease classification using machine learning algorithms. In *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1-5). IEEE.	Dermnet, Dermquest	Colour and texture feature extraction	Decision trees, SVM	Best results: SVM ACC: Colour: 75% Texture: 83%	Texture feature extraction increased accuracy, implemented simple machine learning models	Larger image database necessary for better results, better feature extraction needed
Wu, Z. H. E., Zhao, S., Peng, Y., He, X., Zhao, X., Huang, K., ... & Li, Y. (2019). Studies on different CNN algorithms for face skin disease classification based on clinical images. *IEEE Access*, *7*, 66505-66511.	Xiangya-Derm	-	ResNet50, Inception V3, DenseNet-121, Xception, Inception-ResNet V2	Best results: Inception-ResNet V2 PRE: 70.8% REC: 77.0%	CNN architectures showed overall satisfactory results, pretrained CNNs perform better than non	Precision and recall for AK class of best model low, datasets quality and quantity must be improved

Table 2: Summary of related studies to explainable skin lesion classification

Article Information (APA)	Dataset	Feature representation	Algorithm	Explainability model	Results	Strength	Weakness
Ballari, G. S., Giraddi, S., Chickerur, S., & Kanakareddi, S. (2022). An Explainable AI-Based Skin Disease Detection. In *ICT Infrastructure and Computing: Proceedings of ICT4SD 2022* (pp. 287-295). Singapore: Springer Nature Singapore.	Skin disease dataset from Kaggle	-	ResNet-18	GradCAM	Model accuracy of 96%, Grad-CAM output displays convolutional feature map	High model accuracy, visually interpretable results	Results not clearly communicated by author, dataset not specified, unknown dataset size
Barata, C., Celebi, M. E., & Marques, J. S. (2021). Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, *110*, 107413.	ISIC 2017, ISIC 2018	Colour normalisation	Hierarchical taxonomy method, VGG-16, ResNet-50, DenseNet-161	Trainable attention	Best results on ISIC 2017: VGG-16 SEN: 86.7% SPE: 87.1% AUC: 92.4%	Hierarchical taxonomy bred competitive results, model correctly identifies relevant lesion region, colour normalisation proven to improve accuracy	Melanoma class not easily identified, model not robust to withstand varying transformations on images
Ding, Y., Yi, Z., Li, M., Long, J., Lei, S., Guo, Y., ... & Wang, Y. (2023).	ISIC-2017, ISIC-2018, HAM10000	-	HI-MViT	GradCAM, AblationCAM,	Best results on ISIC 2018 dataset	High results, model can generalize well	Dataset lacks diversity in skin tones and does

HI-MViT: A lightweight model for explainable skin disease classification based on modified MobileViT. *Digital Health*, *9*, 20552076231207197.					PRE: 93.1% REC: 93.2% F1: 93.1% ACC: 93.2%		not other medical indicators
Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., ... & Cha, J. H. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. *Frontiers in Oncology*, *13*, 1151257.	HAM10000, ISIC 2018	-	Xception, Shufflenet, fusion of both models Features selected and fused using IBOA method	GradCAM	Best results HAM10000 dataset, fusion model ACC: 99.3% REC: 99.38% PRE: 99.4% F1: 99.38%	High accuracy for HAM10000 dataset, GradCAM visualization clear explainability model	Bayesian optimization may improve results
Young, K., Booth, G., Simpson, B., Dutton, R., & Shrapnel, S. (2019). Deep neural network or dermatologist?. In *Interpretability of Machine Intelligence in*	HAM10000	-	Inception	GradCAM, KernelSHAP	AUC (mean): 85% REC (mean): 87%	High average AUC and recall across 30 different models	Small dataset size led to spurious correlations by model, Inception-only model limited in accuracy ceiling

Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9 (pp. 48-55). Springer International Publishing.							
Zia Ur Rehman, M., Ahmed, F., Alsuhibany, S. A., Jamal, S. S., Zulfiqar Ali, M., & Ahmad, J. (2022). Classification of skin cancer lesions using explainable deep learning. *Sensors*, *22*(18), 6915.	ISIC	-	MobileNet V2, DenseNet201	GradCAM	Best results: DenseNet201 ACC: 95.50% PRE: 97.02% F1: 95.46% SEN: 93.96% SPE: 97.06%	High scores for metrics	-

Tschandl, P., Argenziano, G., Razmara, M., & Yap, J. (2019). Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. *British Journal of Dermatology*, *181*(1), 155-165.	EDRA, ISIC-2017, private data	-	ResNet50	CBIR	Best results: EDRA dataset ACC: 76.2% AUC: 85.0% SPE: 92.2% SEN: 72.7%	CBIR explainable as content is retrieved based on similarities to other images	Dataset limitations due to lack of diversity, ResNet50 not pretrained for maximum accuracy

Table 3: Summary of related studies to ISIC 2019 classification dataset

Article Information (APA)	Dataset	Feature representation	Algorithm	Explainability model	Results	Strength	Weakness
El-Khatib, H., Popescu, D., & Ichim, L. (2020). Deep learning–based methods for automatic diagnosis of skin lesions. *Sensors*, *20*(6), 1753.	ISIC 2019, PH2	Histogram of oriented gradient	GoogLeNet, ResNet-101, NasNet-Large, decision fusion of all 3 methods	-	ISIC 2019, best results with decision fusion: ACC: 93.00% SPE: 93.33% SEN: 92.50%	Fusion model performed better than individual CNNs, high accuracy, sensitivity and specificity scores	Small dataset used (overall only 300 images)
Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, *7*, 100864.	ISIC 2019	-	EfficientNet	-	Best results with metadata: AUC: 98.00% AUC-S: 96.50% SEN: 55.60% SPE: 99.30%	Including metadata generally increased AUC and specificity scores across all classes, model won ISIC 2019 challenge	AI does not work reliably on out of distribution images, including metadata decreased model sensitivity
Gong, A., Yao, X., & Lin, W. (2020). Dermoscopy image classification based on	ISIC 2019, plus StyleGAN generated images	-	Fusion CNN models	-	DecisionFusion3: ACC: 99.50% AUC: 98.90%	GANs alleviate issues of smaller datasets and imbalanced classes,	Selecting the best combination of fusion CNNs difficult

StyleGANs and decision fusion. *Ieee Access*, *8*, 70640-70650.					PRE: 98.4% SEN: 98.3% SPE: 99.6%	high accuracy for the fusion model	
Hoang, L., Lee, S. H., Lee, E. J., & Kwon, K. R. (2022). Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare. *Applied Sciences*, *12*(5), 2677.	HAM10000, ISIC 2019	-	EW-FCM segmentation, then Wide-ShuffleNet	-	ISIC 2019: ACC: 82.56% SEN: 82.56% SPE: 97.51%	Lightweight model with lesser parameter, higher accuracy than non-segmented models, models do not need ground truth	Performed worse than EfficientNet B0
Iqbal, I., Younus, M., Walayat, K., Kakar, M. U., & Ma, J. (2021). Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized medical imaging and graphics*, *88*, 101843	ISIC 2017, ISIC 2018, ISIC 2019	-	Specialised Deep CNN CSLNet	-	ISIC 2019: ACC: 89.58% PRE: 90.66% SEN: 89.58% SPE: 97.57% F1: 89.75% AUROC: 99.1%	Results consistently outperformed other models for each dataset, multiple kernel units used may aids in recognising symmetry, colour and complex patterns	Further development needed to incorporate age, race, gender information
Kassem, M. A., Hosny, K. M., & Fouad, M. M. (2020). Skin lesions	ISIC 2019		GoogLeNet, GoogLeNet with	-	GoogLeNet ACC: 94.92%	Model outperformed	Multiclass SVM gives lower performance

classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. *IEEE Access*, *8*, 114822-114832.		-	bootstrapped SVM		SEN: 79.8% SPE: 97% PRE: 80.36%	winning ISIC 2019 model	measure on test set
Metta, C., Beretta, A., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S., & Giannotti, F. (2021). Explainable deep image classifiers for skin lesion diagnosis. *arXiv preprint arXiv:2111.11863*.	ISIC 2019	-	ResNet	ABELE	ACC: 0.838 (balanced multiclass)	Saliency maps provide visual representations of explainability	No benchmarks to compare the model against
Nigar, N., Umar, M., Shahzad, M. K., Islam, S., & Abalo, D. (2022). A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, *10*, 113715-113725.	ISIC 2019	-	ResNet-18	LIME	ACC: 94.47% PRE: 93.57% REC: 94.01% F1: 94.45%	Model can generalize well, LIME provides useful visual explanations	Only considers 8 classes of skin disease, dataset does not include opposing examples

Olayah, F., Senan, E. M., Ahmed, I. A., & Awaji, B. (2023). AI techniques of dermoscopy image analysis for the early detection of skin lesions based on combined CNN features. *Diagnostics*, *13*(7), 1314.	ISIC 2019	-	GAC segmentation, then CNN-ANN or CNN-RF models	-	Best model: AlexNet-GoogLeNet-VGG16-ANN ACC: 96.10% SEN: 88.90% SPE: 99.44% PRE: 88.69% AUC: 94.41%	Optimised, segmented and hybrid models achieved high accuracies	
Tô, T. D., Lan, D. T., Nguyen, T. T. H., Nguyen, T. T. N., Nguyen, H. P., & Nguyen, T. Z. (2019). *Ensembled skin cancer classification (ISIC 2019 challenge submission)* (Doctoral dissertation, ISIC2019).	ISIC 2019	-	U-Net segmentation, then EfficientNet-B4	-	-	-	-
Villa-Pulgarin, J. P., Ruales-Torres, A. A., Arias-Garzon, D., Bravo-Ortiz, M. A., Arteaga-Arteaga, H. B., Mora-Rubio, A., ... & Tabares-Soto, R. (2022). Optimized	ISIC 2019, HAM10000	-	DenseNet-201, Inception-V3, Inception-ResNet-V2 (all pretrained on ImageNet)	-	Best model: DenseNet-201 ACC: 93% PRE: 93% REC: 93% F1: 93%	Optimised DenseNet-201 model comparable to SOTA methods	

Convolutional Neural Network Models for Skin Lesion Classification. *Computers, Materials & Continua*, *70*(2).							

Chapter 3: Research Methodology

In this chapter, we introduce the research methodology used for explainable skin lesion classification for this project. There are five subsections in this chapter, in accordance with each step.

3.1 Research design

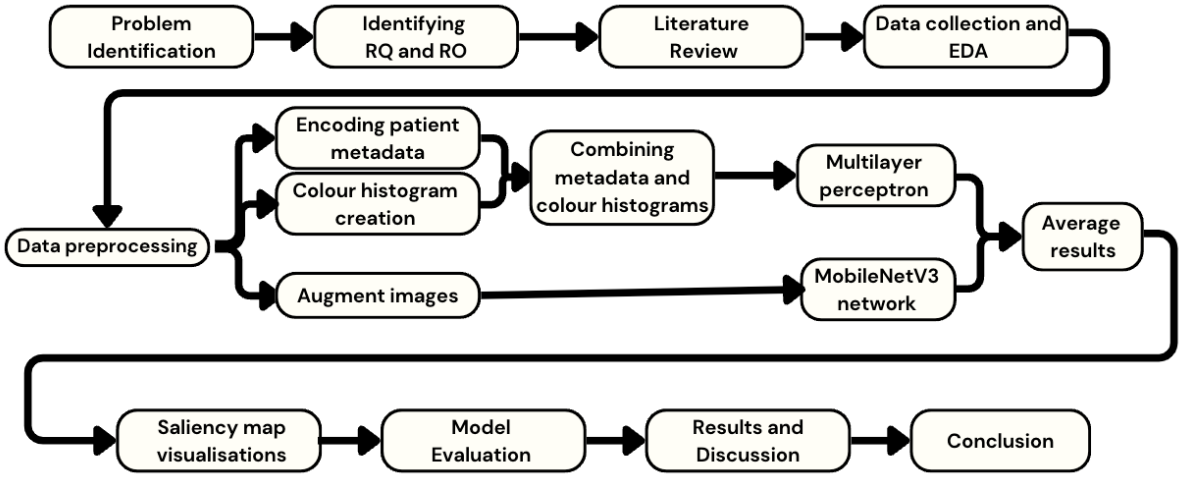


Figure 1 illustrates the research design of the project

For this project, the research design is a follows:

1. Problem identification: identify the research problem and specify the model architecture chosen for experimentation.
2. Data collection: ISIC 2019 dataset from the ISIC Challenge website. The dataset is made up of 3 smaller datasets: HAM10000, ISIC 2017 and BCN20000 (Codella et al., 2018; Combalia et al., 2019; Tschandl et al., 2018).
3. Model development: The model's architecture is as follows:
    a. Colour feature representation: Utilise colour histograms to represent colour features in images.
    b. Integration of metadata: Incorporate patient lesion metadata to increase classification accuracy.
    c. MobileNetV3 architecture: implement MobileNetV3 based architecture for skin lesion classification and analysis
    d. Saliency map visualisations: Include saliency visualisations to highlight relevant parts of the input image and provide model interpretability.
4. Evaluation: Performance metrics utilised will be accuracy, precision, recall, F1-score, sensitivity and specificity.

5. Data analysis: Conduct preliminary exploratory data analysis (EDA) on dataset and analysis on model performance metrics.
6. Results and discussion: Present findings from model performance and explainability mechanisms.
7. Conclusion: Summarise key outcomes, contributions, and highlight areas of improvement for future work.

3.2 Problem Identification

From the literature review in chapter 2, different deep learning models and methods of explainability were analysed and reviewed. From the literature, it is observed:

i. that colour feature representations were underutilised for the ISIC 2019 datasets
ii. only 1 work included patient metadata, despite its availability
iii. saliency maps were an underrepresented explainability method
iv. deep learning methods, such as CNNs lead to the best classification results for precision, recall and F1.

As such, from the literature review, we will be exploring the combination of colour feature representations and patient metadata integration, with MobileNetV3 model architecture and saliency map visualisations for explainability.

3.3 Data collection and preparation

3.3.1 Dataset

The ISIC 2019 dataset from the ISIC Challenge consists of images from 3 smaller datasets: BCN 20000, HAM10000 and the ISIC 2017 Challenge dataset (Codella et al., 2018; Combalia et al., 2019; Tschandl et al., 2018). The dataset contains a total of 25,331 images of skin lesions from 8 different classes of skin lesions: melanocytic nevus (NV), melanoma (MEL), benign kerastosis (BKL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), vascular lesion (VL), dermatofibroma (DF) and actinic kerastosis (AK). Some lesions included are benign, while others such as MEL, BCC and SCC are cancerous. The dataset for this challenge also includes a file on patient metadata, which includes information such as patient age, the site of the lesion, lesion ID and sex.

As the NV class represents a large percentage of skin lesion images in the field, this dataset reflects this fact as the classes are imbalanced. The NV class accounts for over 50% of the dataset, while classes such as VL and DF only have 253 and 239 images respectively. As such, this affects the evaluation metrics chosen for this project.

The distribution of the ISIC 2019 Challenge dataset is as follows:

Class	NV	MEL	BKL	BCC	SCC	VL	DF	AK	TOTAL
Count	12,875	4,522	2,624	3,323	628	253	239	867	25,331

The figure below visualises the different classes of skin lesions within this dataset.

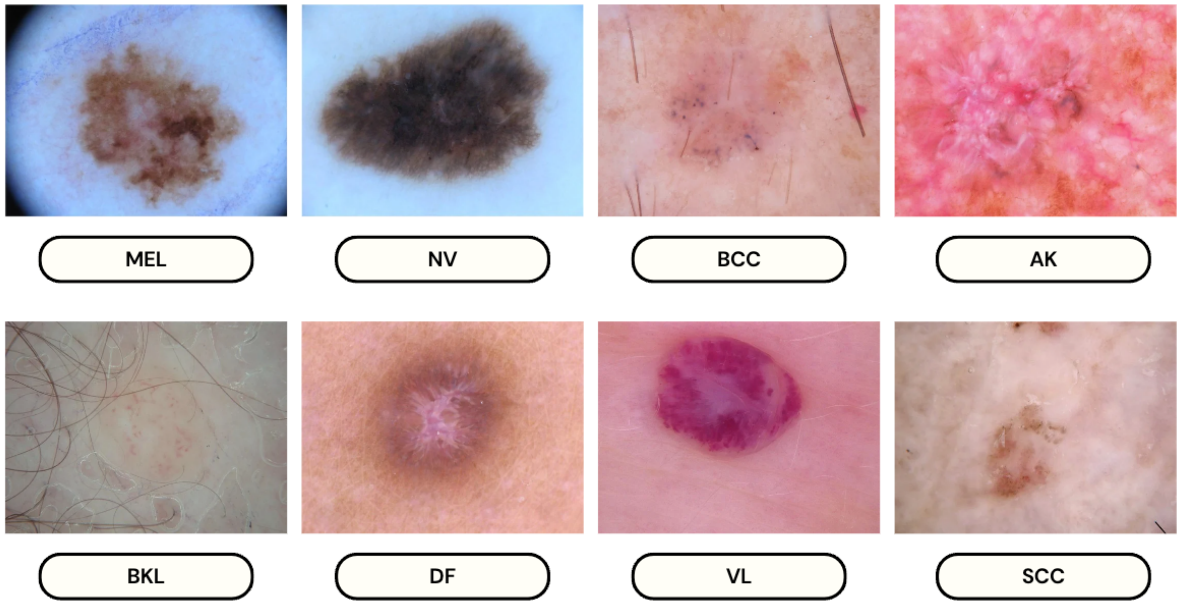


Figure 2: Images of the 8 classes from the ISIC 2019 dataset

3.3.2 Exploratory data analysis (EDA)

In this section, we conduct basic exploratory data analysis on the ISIC 2019 Challenge dataset. The EDA is conducted on the two CSV files associated with the dataset: one contains the image ground truth labels, and the other contains patient metadata. This process allows for thorough examination and understanding of the dataset prior to model building. All EDA is conducted in Python and its associated libraries.

Ground truth dataset

```
df = pd.read_csv('gt.csv')
df.head()
```

	image	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
0	ISIC_0000000	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	ISIC_0000001	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	ISIC_0000002	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	ISIC_0000003	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	ISIC_0000004	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

```
print(df.shape)
```

```
(25331, 10)
```

Analysing the ground truth dataset, we can observe that the dataset contains 25,331 rows and 10 columns. The 10 columns correspond to the image label within the dataset, and the 8 classes of skin lesions. There is an additional column, UNK which is not used in the dataset.

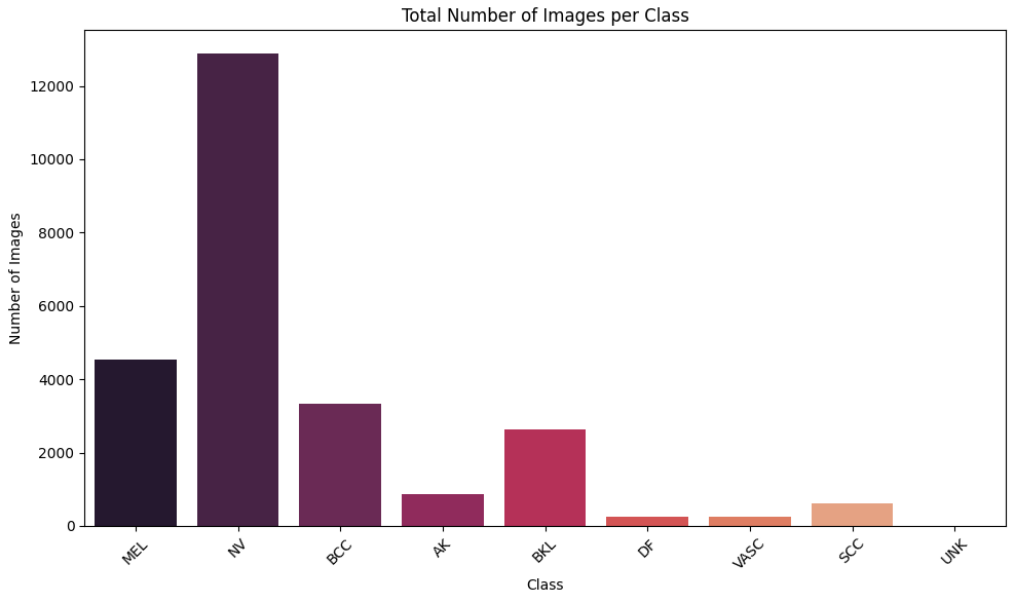


Figure 3: Total number of images per class

From the figure illustrating the total number of images within the dataset, we can observe that the MV class contains the most number of images with over 12,000, followed by the MEL and BCC classes.

Metadata dataset

We analyse the patient metadata in this section.

1. The distribution of the patients' approximate age is left-skewed. The majority of patients are aged 40 years old and above.

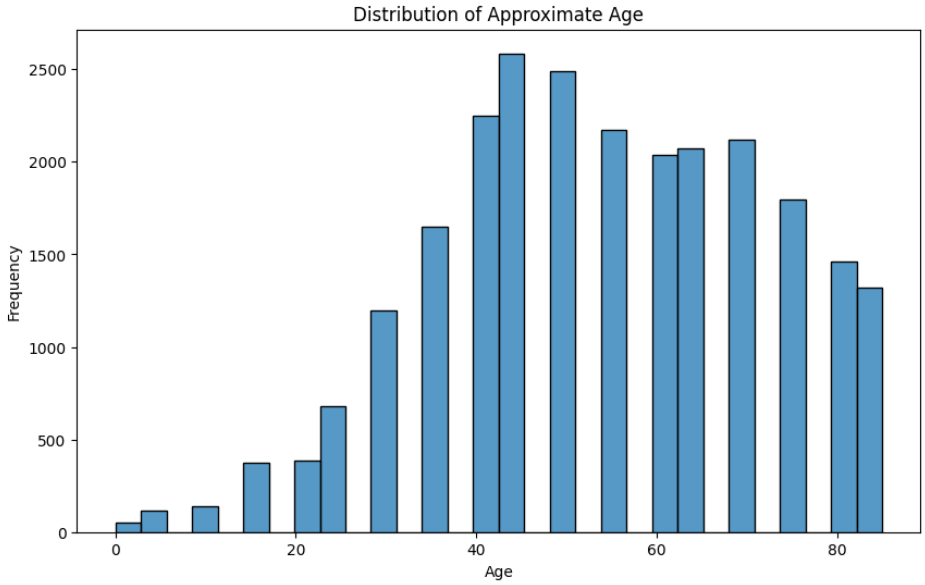


Figure 4: Distribution of approximate patient age

2. In terms of the anatomical sites of skin lesions, the anterior torso is the most common region that skin lesions occur, with nearly 7000 instances. The lower extremity and head/neck area are the next most common regions.

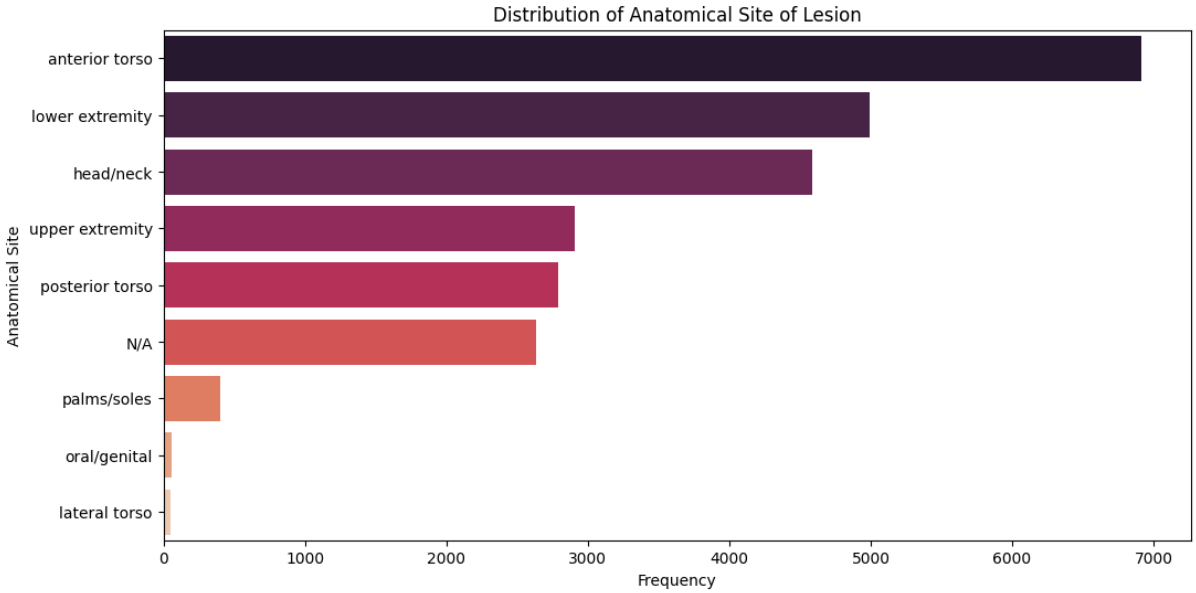


Figure 5: Distribution of Anatomical Site of lesion

3. The number of male and female patients are nearly equal, with 13,286 males and 11,661 female patients.

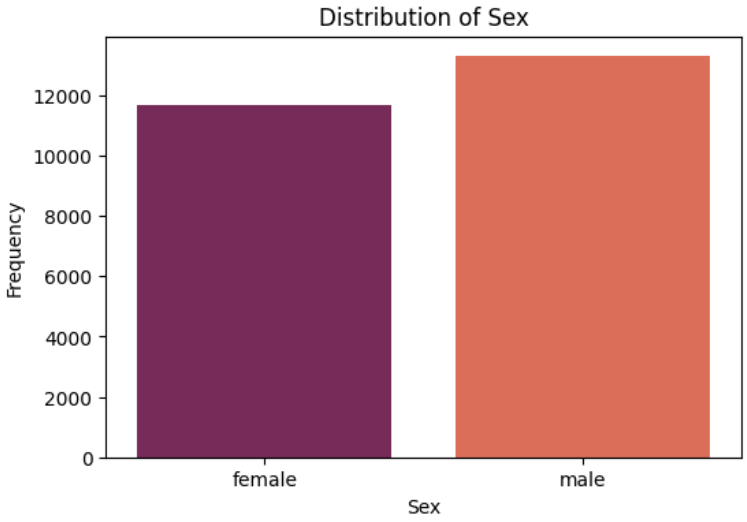


Figure 6: Distribution of patient sex

4. The median age of female patients is approximately 50 years old, while the median age of male patients is around 60 years old. There is a single outlier for the female group and two outliers for the male group.

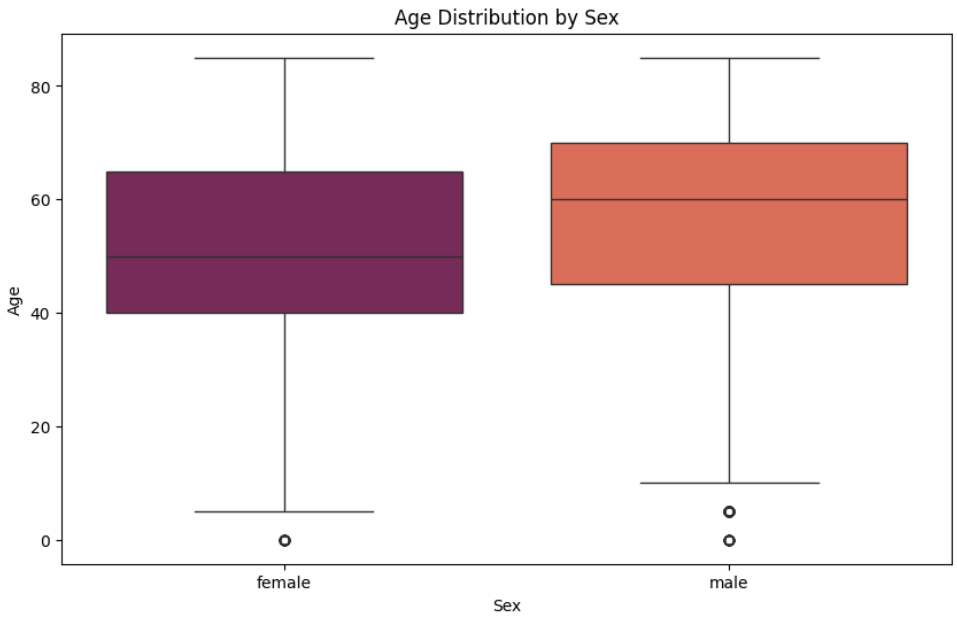


Figure 7: Patient age distribution by sex

3.3.3 Data preprocessing steps

For the ISIC 2019 skin lesion dataset, we aim to implement the following image preprocessing steps:

1. Resizing: Resizing all images to a uniform size suitable for MobileNetV3's architecture. (224 x 224 pixels)
2. Normalisation: Normalise pixel values to a common scale to ensure consistency across the dataset's images.
3. Data Augmentation: Apply data augmentation techniques such as rotation, flipping and scaling to increase dataset diversity and robustness.
4. Colour histograms: RGB histograms are computed and normalized for each channel.
5. Encoding data: Patient metadata, such as age, lesion site, and sex, is encoded using one-hot encoding and normalized to ensure consistency with other features
6. Data splitting: Perform a 70/15/15 split of the data for model training, validation and testing to evaluate the model's performance.

3.4 Proposed model implementation

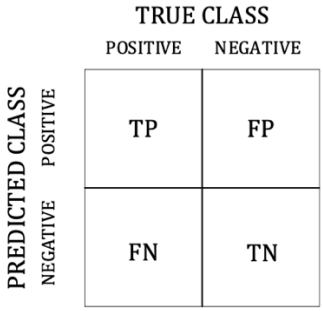
In this section, we introduce the proposed model implementations on the ISIC 2019 dataset. The proposed model aims to improve the accuracy and interpretability of skin lesion classification by integrating colour histograms, patient metadata, and the MobileNetV3 architecture. We detail the comprehensive implementation process, highlighting the model architecture, training procedures, and evaluation metrics.

1. Feature Fusion: Colour histograms and patient metadata are combined into a single feature vector for each image. These combined features are averaged with the image features extracted by MobileNetV3 at the fully connected layer stage, ensuring that both sets of information contribute to the final classification.
2. Model Architecture: The MobileNetV3 (Howard et al., 2019) model is chosen for its efficient and robust feature extraction capabilities. Modifications include additional fully connected layers to process the combined feature vectors from colour histograms and metadata. This hybrid architecture ensures a comprehensive analysis of both image and non-image data.
3. Interpretability and Visualization: Saliency maps are integrated to visualize important regions in the images influencing the model's predictions. This enhances the interpretability of the model and ensures alignment with clinical expectations.

3.5 Model evaluation

This section performs a brief exploration of the project's performance evaluators. For the different models, six machine learning performance indicators are utilised: accuracy, precision, recall, F1-score,

sensitivity and specificity. The figure below illustrates a confusion matrix that will be helpful in understanding the metrics:



Example of a confusion matrix

The confusion matrix's values are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In the context of skin lesion classification, TP indicates **correctly classified skin lesions.** Using these indicators, we calculate the precision, recall and F1 – scores as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision: \frac{TP}{TP + FP}$$

$$Recall: \frac{TP}{TP + FN}$$

$$F1 - score : 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

The results can be interpreted in the following manner:

- **Accuracy:** Number of correct predictions to the total number of predictions. It answers the question: "What percentage of skin lesion images were accurately predicted?"

- **Precision:** Ratio of correctly predicted fake news to predicted fake news. It answers the question, 'For each skin lesion type, what percentage of the images classified as that type were actually of that type?' Precision is used to minimise false positive rates.
- **Recall:** Ratio of correctly predicted fake news to all the observations in the fake news class. It answers the question, 'For each skin lesion type, what percentage of the actual images of that type were correctly identified by the model?' We use recall to minimise false negative rates.
- **F1 – score:** Weighted average of precision and recall. It poses the question, "How well does the model balance precision and recall for each skin lesion type?"
- **Sensitivity:** How well a machine learning model can detect positive instances. The question posed is the same as recall.
- **Specificity:** How well a machine learning model can detect negative instances. It answers the question "For each skin lesion type, what percentage of the images that are not of that type were correctly identified as not being that type?"

Chapter 4: Implementation

In this chapter, we elaborate the implementation details of our experiments. Six sections are outlined, detailing the preprocessing, augmentation, colour histogram creation, model architectures, saliency maps and experimental setup steps.

4.1 Preprocessing

As our proposed model contains both image and metadata, we preprocess both data configurations, with the details highlighted in the following section.

4.1.1 Metadata preprocessing

For our metadata, we employ the following preprocessing steps:

1. Load the dataset
2. Remove unnecessary class UNK
3. Handled missing values from the 'age_approx', 'anatom_site_general', 'sex' with median and mode fill replacements.
4. One-hot encode the 'anatom_site_general' and 'sex' columns
5. After preprocessing, the metadata dataset had shape (25331, 12).
6. Stratify split the dataset into train, validation and test datasets based on a 70/15/15 split. This ensures that the ratio of each class is maintained in all datasets. The names of the datasets were X_train, X_val, X_test, y_train, y_val and y_test respectively. The train dataset contained 17,331 instances, while the testing and validation datasets contained 3,800 instances each.

4.1.2 Image preprocessing

For our images, we employed minimal preprocessing steps as below:

1. Create train, validation and test directories.
2. Within each directory, create 8 folders representing the 8 classes.
3. Copy the images into the train, validation and test directories in accordance to their label. The images are split based on their 'image' id in the X_train, X_val and X_test metadata datasets.

4.2 Data augmentation

As the original dataset is highly imbalanced, we perform data augmentation to balance the classes better and prevent overfitting. This is done for both the image and metadata data respectively.

4.2.1 Image data augmentation

We augment the images for the training dataset only to prevent training overfitting and misclassification. For the training images, we perform the following:

1. Set a target count of 3000. This number was chosen as it was near the count of the second highest class NV, which had 3165 images.
2. Create an ImageDataGenerator that defines the augmentation configuration. For our experiments, we set the following configurations:
    - Rotation_range = 20
    - Width_shift_range = 0.2
    - Height_shift_range = 0.2
    - Shear_range = 0.2
    - Zoom_range = 0.2
    - Horizontal_flip = True
    - Vertical_flip = True
    - Fill_mode = “nearest”
    - Brightness_range = [0.8, 1.2]
3. For every class within the train directory, calculate the number of images in that class.
    a. If the number of images was less than the target count, create augmented images using the ImageDataGenerator until it reached that target count.
    b. For images exceeding the target count, do not create any augmented images.
4. After augmentation, the train dataset count went from 17,331 images to 30,177 images. Class count after augmentation:

Table 4: Count of images before and after augmentation

	MEL	***NV***	***AK***	***BCC***	***BKL***	***DF***	***SCC***	***VASC***
Count before augmentation	3165	9012	607	2326	1837	167	440	177
Count after augmentation	3165	9012	3000	3000	3000	3000	3000	3000

4.2.2 Metadata augmentation

After performing image augmentation, the metadata was augmented as well. A dataframe was created which included the information of the original images, and the metadata of the augmented images,

which were copied based on the original image id. The shape of the X_train was (30177, 11) after augmentation and removal of the 'image' column.

4.3 Colour histogram

To include colour histograms in our datasets, we do the following:

1. Compute colour histograms for each image within the train, validation and test datasets. The bin size chosen was 8, and with 3 colour channels, 256 values were computed. These histograms provided a high-level colour histogram overview of the images. The pixel values were then normalised.
2. The histogram values were then appended to the train, test and validation datasets respectively. For X_train, the shape after appending the values is (30177, 523)

4.4 Model architecture

In this section, we outline the architecture of the models used for the experiment. Two models were created: MobileNetV3 and MLP + MobilenetV3

4.4.1 MobileNetV3 model

For our experiments, we chose the MobileNetV3Small model pretrained on the ImageNet dataset. The model has approximately 2.5 million parameters , with a size of 4MB. After passing through the base MobileNetV3Small model, the input goes through a GlobalAveragePooling2D layer, a Fully Connected (FC) layer with 128 units and ReLu activation, before passing through the classification layer with softmax activation. Only images are passed through the MobileNetV3 model.

4.4.2 Multilayer perceptron model (MLP)

To classify the metadata, a 3-layer MLP was created to learn from the features. It consists of two FC layers with ReLu activation and 128 and 64 units respectively, and a classification layer with softmax activation to classify the output.

4.4.3 Ensemble model (MLP + MobileNetV3)

This model classifies both the metadata and images. Firstly, the image data and metadata are trained separately by the MobileNetV3Small model and MLP model respectively. The outputs from both models are then averaged with an Average() layer, with both models contributing equally to the final prediction.

4.5 Saliency maps

For each image, we generate saliency maps to identify which parts of an image are most influential in the model's decision-making process. It calculates the gradients of this prediction with respect to the input image. The saliency map is created by taking the average of these gradients across the different colour channels, highlighting the parts of the image that most affect the model's prediction.

4.6 Experimental setup

The experimental setup is as follows for both models:

- Training, validation and test generators were created for each set of data. A custom data generator was created for the ensemble model to include both image and meta data, while the MobileNetV3 only model utilised Keras' ImageDataGenerator function.
- The MobileNetV3 only
- Batch size is set at 32
- Images are resized to the shape (224, 224)
- Learning rate is set at 0.0001
- Compiling the model, Adam optimiser is used, categorical cross entropy loss calculated and accuracy being the metric to track
- Training is set at 50 epochs.
- EarlyStopping is included to prevent model overfit for the validation loss.

Chapter 5: Results and Discussion

In this chapter, we discuss the results of our models classifying skin lesions into 8 different classes. We first introduce the baseline model our results are compared to, followed by the results of our two models. We then discuss the results, and offer interpretations. Finally, we will end the chapter discussing the limitations of our work.

5.1 Baseline

As a baseline, we use the results achieved by the winner of the ISIC 2019 Challenge (Gessert et al., 2020) as a benchmark against our findings. The authors won the Task 2 challenge, which incorporates both the metadata and the images within the dataset. For the metadata, they used an MLP model, while implementing EfficientNet models as their CNN component. Their best model accuracy was 63% for the models that both include and exclude metadata, with a mean sensitivity of 73% for the model that included metadata.

5.2 Results

In this section, we visualise the results obtained from the two models used. As mentioned in the previous chapter, all models were train with a batch size of 32 and with 50 epochs. For each model, we display the per-class accuracy, precision, recall, F1-score, sensitivity and specificity. We also outline the overall accuracy, macro-averaged and weighted averaged precision, recall and F1-scores.

5.2.1 MobileNetV3 – Image only (Model 1)

Table 5: Metric results achieved by the MobileNetV3 only model

Class	**Metrics**					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Sensitivity**	**Specificity**
MEL	7.37%	15.77%	7.37%	10.05%	91.45%	7.37%
NV	**78.16%**	**50.40%**	**78.16%**	**61.28%**	20.45%	**78.16%**
BCC	0.60%	6.25%	0.60%	1.10%	98.64%	0.60%
AK	10.00%	5.06%	10.00%	6.72%	93.35%	10.00%
BKL	1.27%	13.16%	1.27%	2.31%	99.03%	1.27%
DF	0.00%	0.00%	0.00%	0.00%	**99.39%**	0.00%
VASC	5.26%	3.08%	5.26%	3.88%	98.33%	5.26%
SCC	2.13%	3.57%	21.28%	2.67%	98.54%	2.13%

From Table 5, we can observe that the NV class of images achieved the highest accuracy, precision, recall, F1-scores and specificity results. Meanwhile, the highest sensitivity scores belonged to the DF

class, which scored 0% on all other metrics, indicating that no image from the DF class was accurately predicted.

Table 6: Accuracy, macro-averaged and weighted-average precision, recall and F1-scores for the MobileNetV3 only model

	Accuracy	Precision	Recall	F1-Score
Accuracy	41.71%			
Macro avg		12.16%	13.10%	11.00%
Weighted avg		30.91%	41.71%	33.67%

The test dataset for the image-only model achieved an overall classification accuracy of 41.71%. Meanwhile, the macro-averaged precision, recall and F1-scores all fall below 14%. The weighted-average scores are slightly better at 30.91%, 41.71% and 33.67% respectively for the three metrics; however they all fall below the 50% threshold.

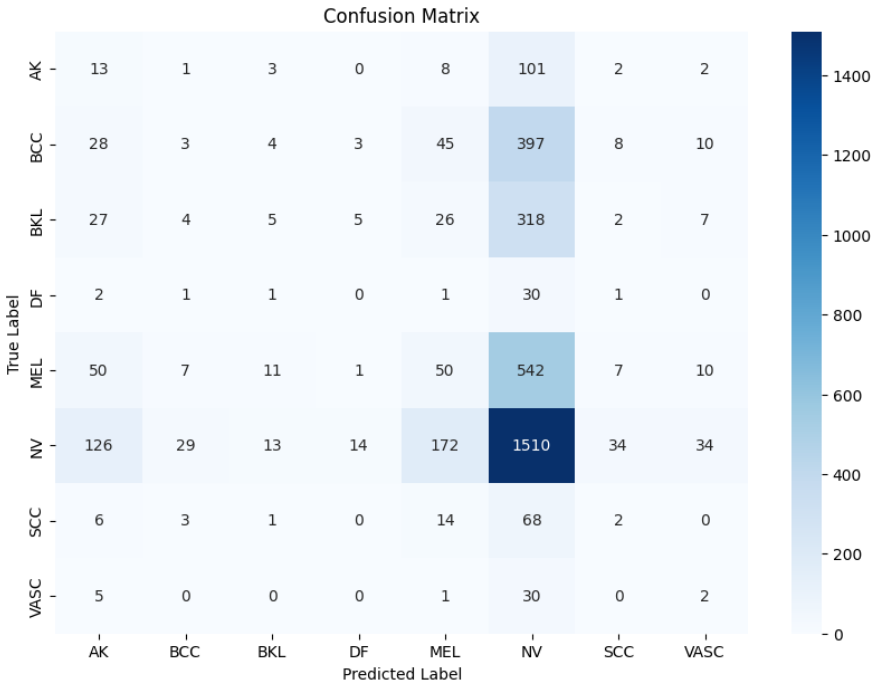


Figure 8: Confusion matrix of the MobileNetV3 only model results

From the confusion matrix for this model, we can observe that:

- NV class of images was classified correctly the most, with 1510 correctly predicted images, as compared to the other classes.
- Furthermore, the confusion matrix illustrates that most other images were predicted to come from the NV class as well. This can be attributed to the size of the NV class in the dataset, as it represents over 50% of the images in the test set.
- Besides the NV class, the MEL and AK classes round up the most-predicted labels.

- No images from the DF class were correctly predicted during testing.

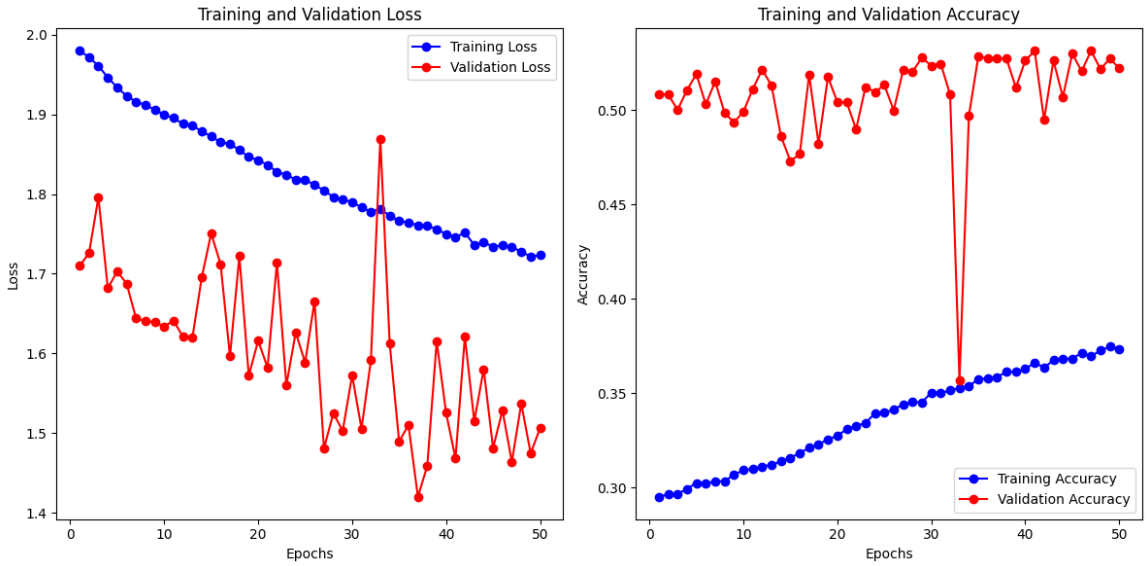


Figure 9: Training and validation loss (left) and accuracy (right) of the MobileNetV3 only results over 50 epochs

From Figure 9, we can observe a few key points for the training and validation loss and accuracies respectively:

- In the loss graph, the training loss is higher than the validation loss. Generally, during the opposite is true, ie the validation loss is higher than the training loss.
- Likewise, a similar issue is spotted in the accuracy graph. The training accuracy is lesser than the validation accuracy, when generally the inverse is true. The reasons for this discrepancy will be discussed in the next section.
- Generally, both the training and validation losses are steadily decreasing at similar rates. The training and validation accuracies are increasing as well with the epochs; however, training accuracy is increasing at a higher rate than validation accuracy.

5.2.2 MLP-MobileNetV3 – Image and metadata (Model 2)

Table 7: Metric results achieved by the MLP-MobileNetV3 model

Class	Metrics					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Sensitivity**	**Specificity**
MEL	7.10%	84.21%	7.10%	13.09%	99.71%	7.10%
NV	56.83%	**86.50%**	56.83%	**68.60%**	90.85%	56.83%
BCC	8.67%	40.95%	8.67%	14.31%	98.11%	8.67%
AK	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
BKL	**83.03%**	13.77%	**83.03%**	23.63%	40.30%	83.03%
DF	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
VASC	2.63%	11.11%	2.63%	4.26%	**99.79%**	**92.63%**
SCC	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

From Table 7, we observe that the BKL class achieves the highest accuracy and recall scores at 83.03% for both. Meanwhile, the NV class receives the highest precision and F1-scores at 86.50% and 68.60% respectively. The AK, DF, and SCC classes all have a specificity score of 0% and 0% across all other metrics, indicating no images were correctly classified for those labels. Finally, the VASC class achieves a highest specificity of 92.63%.

Table 8: Accuracy, macro-averaged and weighted-average precision, recall and F1-scores for the MLP-MobileNetV3 model

	Accuracy	**Precision**	**Recall**	**F1-Score**
Accuracy	40.00%			
Macro avg		30.00%	20.00%	15.00%
Weighted avg		66.00%	40.00%	42.00%

From the table above, we note that the model's overall accuracy is 40%, while its macro-averaged precision, recall and F1-scores are 30%, 20% and 15% respectively. The weighted average results fare better at 66%, 40% and 42% for the three metrics.

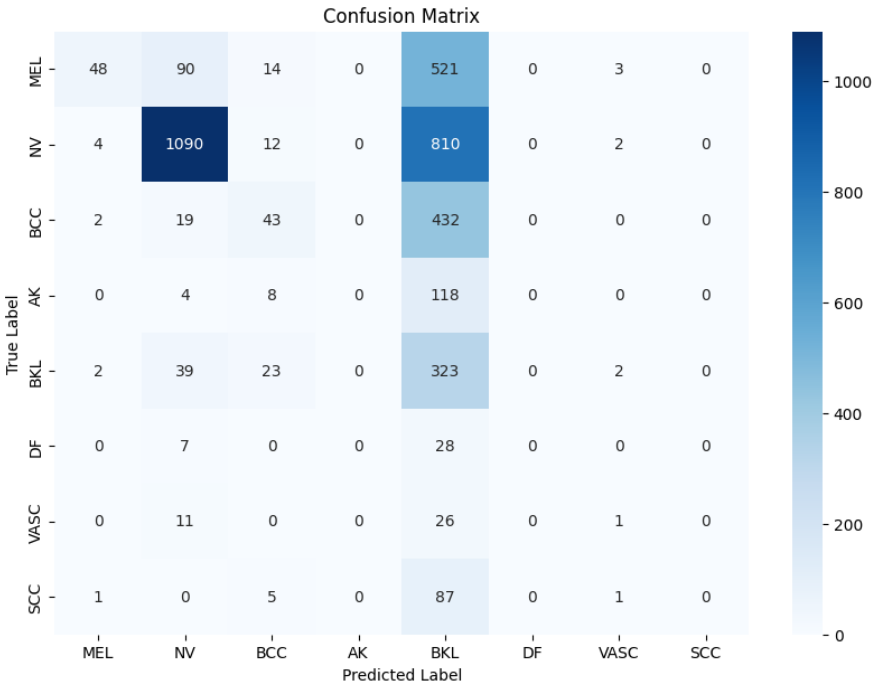


Figure 10: Confusion matrix of the MLP-MobileNetV3 model results

From the confusion matrix of the MLP-MobileNetV3 model, we can make the following observations:

- The NV class was the most correctly predicted class, with 1090 correctly predicted images.
- The BKL class was the most predicted label, as all class labels predicted their images to belong to this class.
- After the BKL class, most images were predicted to belong to either NV or BCC.
- There were no correctly predicted images in the AK, DF and SCC classes. Most of their predicted labels belonged to the BKL class.

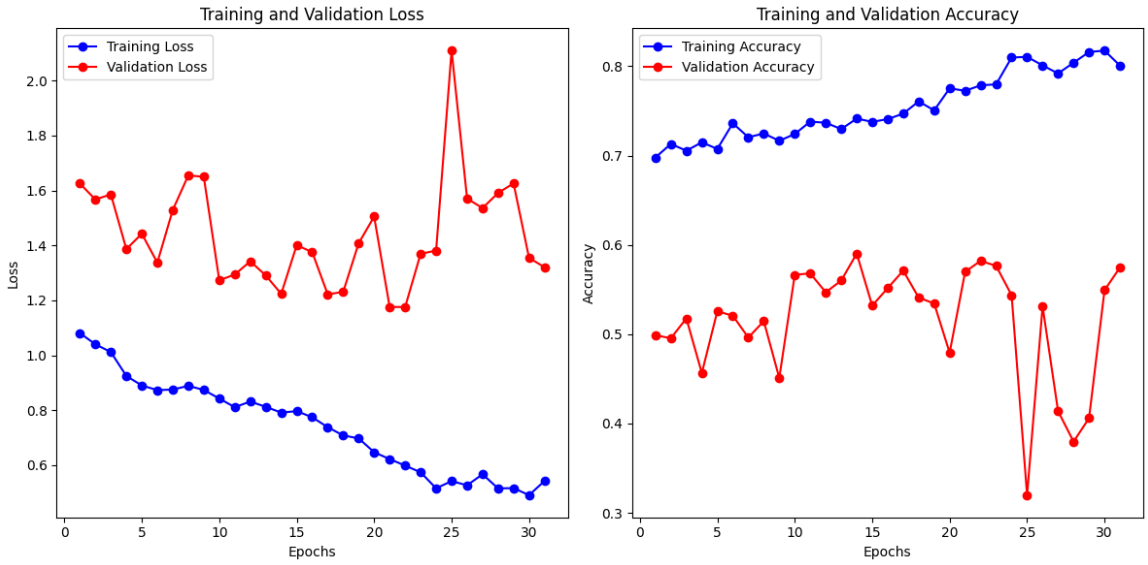


Figure 11: Training and validation loss (left) and accuracy (right) of the MLP-MobileNetV3 results over 31 epochs

From the figure above illustrating the training and validation losses and accuracies, we can infer the following:

- The training and validation losses follow the general trend, ie that training loss is well below the validation loss. Similarly, the training accuracy plots higher than the validation accuracy, which is the conventional trend.
- Generally, both the training and validation losses decrease with the epochs. The inverse is true for the train and validation accuracies, where both plots are increasing over epochs.
- The training stops at 31 epochs. This is due to the early stopping callback implemented as detailed in the previous chapter.

5.3 Discussion

In this section, we discuss the results of our augmented MobileNetV3 model, and our augmented, colour histogram, metadata-included MLP-MobileNetV3 hybrid model for skin lesion classification.

In terms of overall model accuracy, the image-only model achieved the higher accuracy 41.71%, as compared to the hybrid model's 40% classification accuracy. Meanwhile, the hybrid model achieved higher metric scores overall on their data, with their accuracy, precision, recall and F1-scores better than those achieved of by the image-only model. Despite the MobileNetV3 model having the slightly higher accuracy, realistically this model would not make it to deployment as the classification accuracy is too low for medical standards. Accurate skin lesion classifications save lives; models with subpar accuracy scores will not be deployed, as they can risk human lives and company reputations.

Comparing the weighted average metrics for both models, which adjust for class imbalance in the dataset, we find that the MLP-MobileNetV3 model outperforms the MobileNetV3 model in terms of precision by 35 percentage points. This indicates it's better at reducing false positives, especially for frequent classes. However, the MobileNetV3 model has a slightly higher recall (41.71% vs. 40%), suggesting it's better at identifying all relevant instances and minimizing false negatives, which is crucial for tasks like detecting skin lesions. The MLP-MobileNetV3 model also has a higher weighted-average F1-score (42%), indicating a better balance between precision and recall, making it more robust overall.

Upon closer examination of the confusion matrices of the MobileNetV3 and MLP-MobileNetV3 models, we can observe that the first model could classify the images into all 8 classes, regardless of true labels. Meanwhile, the latter model did not predict any images to belong to the AK, DF and SCC classes, despite the AK class containing over 120 images. This could be because the latter model stopped training due to the EarlyStopping implementation; the patience of the callback was set at 5,

which in hindsight can be considered too low leading to the model not being able to train on all the classes sufficiently. As such, the MLP-MobileNetV3 model was not able to identify and classify the AK, DF and SCC confidently. Future work should look into setting a higher patience level, or allowing the model to train for the full 50 epochs.

Compared to the MobileNetV3 only model, the MLP-MobileNetV3 model had a lower train loss and higher train accuracy scores from the initial run. Furthermore, the validation loss for the latter model kept steadily decreasing with each passing epoch, and would have likely been much lower if the MLP-MobileNetV3 model had trained for the entire 50 epochs instead of the 31 epochs. While it cannot be stated conclusively, this can suggest that the incorporation of metadata can lead to better image classification scores for skin lesions.

5.4 Comparison against baseline

We now turn our attention to comparing our results with the established baseline set by Gessert et al. (2020). The MobileNetV3-only model in our study achieved an accuracy of 41.71%, which falls significantly short of the 63% accuracy reported by the authors. Additionally, our MLP-MobileNetV3 hybrid model, while incorporating additional features, reached a maximum accuracy of 40% and a mean sensitivity of 54%. These figures are considerably lower than the 63% accuracy and 73% sensitivity achieved by the baseline.

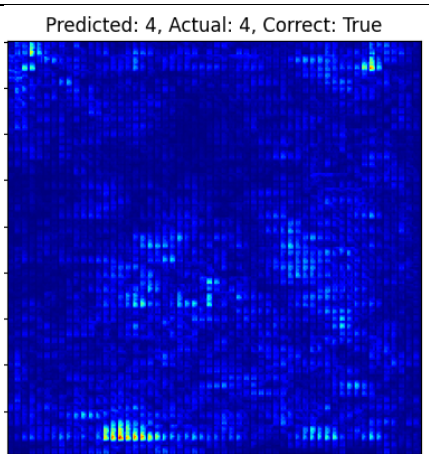
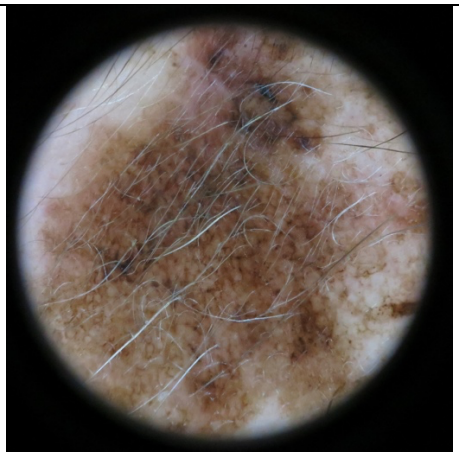
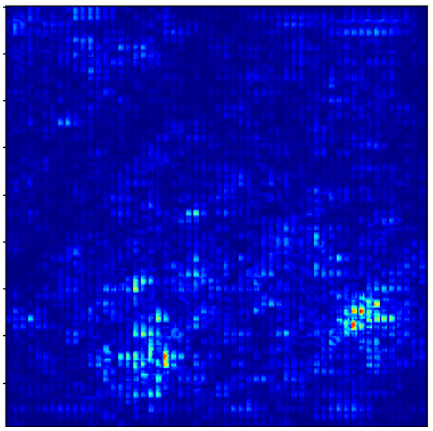
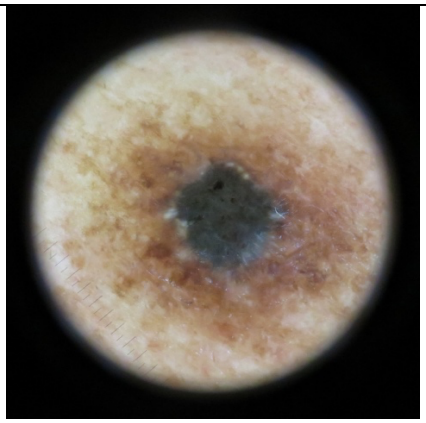
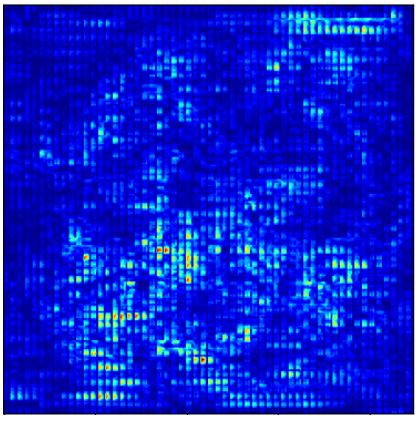
The significant disparity in performance between our models and the benchmark underscores the inherent challenges and complexities involved in this research domain. Despite efforts to enhance model architecture and integrate multiple learning approaches, our results suggest that there may be underlying factors, such as data quality, preprocessing techniques, or specific model configurations, that contribute to the performance gap.

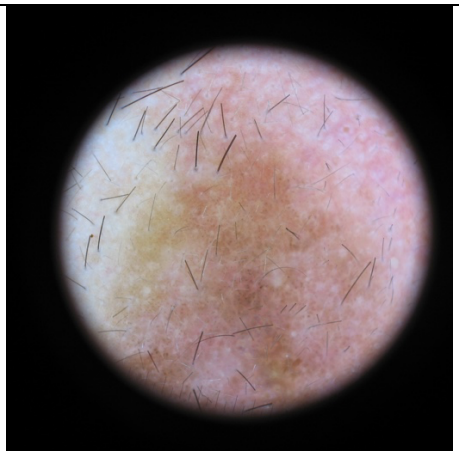
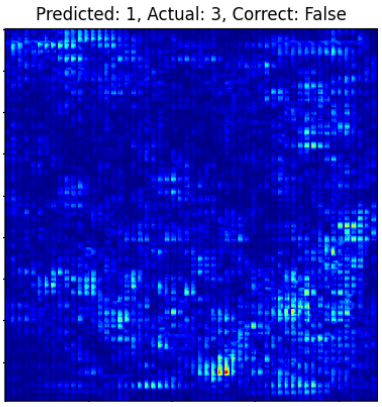
While these findings highlight areas for further improvement, they also offer valuable insights into the limitations and potential areas for refinement in future work. The comparison against Gessert et al. (2020) serves as a critical benchmark, reminding us of the progress yet to be made in this rapidly evolving field.

5.5 Saliency maps interpretation

In this section, we visualise and interpret a few saliency maps created after running the image through the MLP-MobileNetV3 model. We contrast them against the original images to better understand the

explainable model. Below, we visualise the images and saliency maps of the images IDs ISIC_0072553, ISIC_0072601, ISIC_0065647, ISIC_0072999.

	Original Image	Saliency map
ISIC_0072773		
ISIC_0072601		
ISIC_0065647		

ISIC_0072999		

For the **correctly predicted images**, such as images ISIC_0072773 and ISIC_0072601, the saliency maps reveal that the model predominantly focuses on the relevant features associated with the correct class. The lighter areas in these saliency maps highlight the parts of the image that significantly influenced the model's accurate classification. This suggests that the model is effectively identifying and leveraging the key features that correspond to the target class.

Conversely, for **wrongly predicted images**, such as images ISIC_0065647 and ISIC_0072999, the saliency maps indicate that the model's focus is misplaced. The lighter regions in these maps show where the model concentrated its attention, but these areas may not align with the relevant features of the true class. This misalignment suggests that the model is potentially relying on irrelevant or misleading features, leading to the incorrect prediction.

Overall, while lighter colours in saliency maps illustrate areas of high influence for the model's predictions, interpreting these areas in the context of prediction accuracy provides insights into whether the model is attending to the correct or incorrect features of the image.

5.6 Limitations

Below, we outline several key limitations that should be considered to enhance future work. Addressing these challenges will enable us to refine training protocols, improve data management strategies, and boost model interpretability, ultimately maximizing the potential of the MLP-MobileNetV3 architecture.

5.6.1 Computational demands

Both the MobileNetV3 and MLP-MobileNetV3 models, while designed for efficiency, still face computational limitations, particularly with extensive training. In our experiments, we observed that the MobileNetV3 model could benefit from additional training beyond 50 epochs to fully converge and optimize performance, as the validation loss is still decreasing at the 50 epoch mark. We assume the same for the MLP-MobileNetV3 model, as it did not reach 50 epochs during training. Training for a longer duration could potentially improve accuracy and robustness, given the model's complex architecture and integration of multiple components.

5.6.2 Data dependency

The performance of the MLP-MobileNetV3 model is highly contingent on the quality and quantity of the training data. The imbalanced nature of the dataset can adversely affect model performance, leading to suboptimal results. Adequate data preprocessing and augmentation strategies are essential to mitigate this limitation and enhance the model's generalization capabilities.

5.6.3 Interpretability

While saliency maps provide valuable insights into which regions of an image influence the model's predictions, interpreting these maps can be challenging. In cases of incorrect predictions, the saliency maps may reveal that the model is focusing on irrelevant features, which complicates the diagnosis of the model's decision-making process. This lack of clarity can hinder efforts to understand and refine model behavior.

5.6.4 Resource constraints

Despite the efficiency improvements of MobileNetV3, the combined MLP-MobileNetV3 architecture still demands substantial computational resources during training and inference. This can be a limiting factor in environments with constrained computational power or when scaling up to larger datasets.

5.6.5 Model complexity

The integration of MLP with MobileNetV3 adds to the model's complexity, which can affect both training time and inference speed. Optimizing hyperparameters and ensuring efficient model architecture are crucial to balancing performance and computational efficiency.

Chapter 6: Conclusion

In this research report, we sought out to create an explainable skin lesion classification model by investigating the existing literature for the best deep learning architecture, feature representations and explainability model. From there, we analysed the information and proposed that an MLP-MobileNetV3 model that incorporated patient metadata and colour histogram values would be the most fruitful in obtaining high skin lesion classification results.

Experimenting on the ISIC 2019 dataset which include patient metadata, our findings demonstrate that while the MLP-MobileNetV3 model shows promise, particularly in terms of precision and sensitivity, it does not excel in prediction accuracy. The model's performance, though underperforming compared to the baseline set by Gessert et al. (2020)'s, with our 40% vs their 63% in accuracy and 54% vs 73% in sensitivity, provides valuable insights into the integration of metadata and advanced architectures.

The use of saliency maps has been instrumental in interpreting the model's decision-making process. By visualizing which regions of an image influence predictions, we gained a better understanding of the model's focus, both in accurate and incorrect predictions. This interpretability is crucial for developing trustworthy AI systems, particularly in sensitive applications like skin lesion classification, where understanding model decisions can significantly impact clinical outcomes and patient trust.

Accurate classification of skin lesions is critical due to its direct implications for patient health and treatment decisions. Although our model has room for improvement, it represents a step toward enhancing diagnostic tools with advanced machine learning techniques. Future research should address the model's limitations, such as extending training durations and exploring alternative architectures. Enhancing the model's interpretability and reliability through further refinement of saliency maps and other explainability tools will also be vital. By tackling these challenges, we can advance the effectiveness of image classification models and their practical applications in healthcare. Overall, this study provides a foundation for future advancements in accurate and interpretable skin lesion classification systems.

Chapter 7: Reference List

Ahmad, B., Usama, M., Huang, C.-M., Hwang, K., Hossain, M. S., & Muhammad, G. (2020). Discriminative feature learning for skin disease classification using deep convolutional neural network. *IEEE Access*, *8*, 39025-39033.

Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., Kim, Y. J., & Cha, J.-H. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. *Frontiers in oncology*, *13*, 1151257.

Allugunti, V. R. (2022). A machine learning model for skin disease classification using convolution neural network. *International Journal of Computing, Programming and Database Management*, *3*(1), 141-147.

Ballari, G. S., Giraddi, S., Chickerur, S., & Kanakareddi, S. (2022). An Explainable AI-Based Skin Disease Detection. In *ICT Infrastructure and Computing: Proceedings of ICT4SD 2022* (pp. 287-295). Springer.

Barata, C., Celebi, M. E., & Marques, J. S. (2021). Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, *110*, 107413.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., & Kittler, H. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018),

Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A. C., & Puig, S. (2019). Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

Ding, Y., Yi, Z., Li, M., Long, J., Lei, S., Guo, Y., Fan, P., Zuo, C., & Wang, Y. (2023). HI-MViT: A lightweight model for explainable skin disease classification based on modified MobileViT. *Digital Health*, *9*, 20552076231207197.

El-Khatib, H., Popescu, D., & Ichim, L. (2020). Deep learning–based methods for automatic diagnosis of skin lesions. *Sensors*, *20*(6), 1753.

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, *7*, 100864.

Gong, A., Yao, X., & Lin, W. (2020). Dermoscopy image classification based on StyleGANs and decision fusion. *IEEE Access*, *8*, 70640-70650.

Hoang, L., Lee, S.-H., Lee, E.-J., & Kwon, K.-R. (2022). Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare. *Applied Sciences*, *12*(5), 2677.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., & Vasudevan, V. (2019). Searching for mobilenetv3. Proceedings of the IEEE/CVF international conference on computer vision,

Iqbal, I., Younus, M., Walayat, K., Kakar, M. U., & Ma, J. (2021). Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized medical imaging and graphics*, *88*, 101843.

Kassem, M. A., Hosny, K. M., & Fouad, M. M. (2020). Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. *IEEE Access*, *8*, 114822-114832.

Kurasinski, L., & Mihailescu, R.-C. (2020). Towards machine learning explainability in text classification for fake news detection. 2020 19th IEEE international conference on machine learning and applications (ICMLA),

Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Metta, C., Beretta, A., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S., & Giannotti, F. (2021). Explainable deep image classifiers for skin lesion diagnosis. *arXiv preprint arXiv:2111.11863*.

Ni, S., Li, J., & Kao, H.-Y. (2021). MVAN: Multi-view attention networks for fake news detection on social media. *IEEE Access*, *9*, 106907-106917.

Nigar, N., Umar, M., Shahzad, M. K., Islam, S., & Abalo, D. (2022). A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, *10*, 113715-113725.

Olayah, F., Senan, E. M., Ahmed, I. A., & Awaji, B. (2023). AI techniques of dermoscopy image analysis for the early detection of skin lesions based on combined CNN features. *Diagnostics*, *13*(7), 1314.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Swamy, K. V., & Divya, B. (2021). Skin disease classification using machine learning algorithms. 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4),

Tô, T. D., Lan, D. T., Nguyen, T. T. H., Nguyen, T. T. N., Nguyen, H.-P., & Nguyen, T. Z. (2019). *Ensembled skin cancer classification (ISIC 2019 challenge submission)* ISIC2019].

Tschandl, P., Argenziano, G., Razmara, M., & Yap, J. (2019). Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. *British Journal of Dermatology*, *181*(1), 155-165.

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

Villa-Pulgarin, J. P., Ruales-Torres, A. A., Arias-Garzon, D., Bravo-Ortiz, M. A., Arteaga-Arteaga, H. B., Mora-Rubio, A., Alzate-Grisales, J. A., Mercado-Ruiz, E., Hassaballah, M., & Orozco-Arias, S. (2022). Optimized convolutional neural network models for skin lesion classification. *Computers, Materials & Continua*, *70*(2).

Wu, Z., Zhao, S., Peng, Y., He, X., Zhao, X., Huang, K., Wu, X., Fan, W., Li, F., & Chen, M. (2019). Studies on different CNN algorithms for face skin disease classification based on clinical images. *IEEE Access*, *7*, 66505-66511.

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., & Hu, X. (2019). Xfake: Explainable fake news detector with visualizations. The world wide web conference,

Young, K., Booth, G., Simpson, B., Dutton, R., & Shrapnel, S. (2019). Deep neural network or dermatologist? Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9,

Zia Ur Rehman, M., Ahmed, F., Alsuhibany, S. A., Jamal, S. S., Zulfiqar Ali, M., & Ahmad, J. (2022). Classification of skin cancer lesions using explainable deep learning. *Sensors*, *22*(18), 6915.