King's College London

Quantifying Explainable AI Methods in Medical Diagnosis: A study in skin cancer

Submitted by:
Hardik Sangwan

Project Report Submitted in Partial Fulfillment of the Requirements of the MSc Project Module

Engineering Department

Supervised by:
Prof. Bipin Rajendran

London
August/2024

Abstract

Deep Learning has shown extensive promise, in academic research, for diagnosing disease. It can improve patient outcomes while reducing professional workloads. But these diagnostic methods are still not prevalent in clinical settings, and rightfully so, due to issues surrounding trustworthiness and explainability. This thesis presents a comprehensive new study on quantifiably measuring explainable AI techniques, in the context of diagnosing skin cancer, to increase trustworthiness and model adoption in real-life settings.

Firstly, classification models were trained to achieve state-of-the-art metrics (precision, recall, accuracy) using the HAM10000 dataset containing seven diagnostic categories related to skin cancer. The models included Convolutional Neural Networks (CNNs) built from scratch along with architectures such as DenseNet, MobileNet, and ResNet trained through transfer learning. The dataset was split into training, validation, and test subsets, with augmentation, oversampling and hyperparameter tuning techniques applied to enhance global test accuracy.

Once the models had been trained, methods such as SHAP, LIME, and Integrated Gradients were utilized to generate feature-based explanations for model predictions. The final step, wherein lies the novel contribution of this thesis, was the quantification of these visual explanations. Metrics such as faithfulness, robustness and complexity were used to evaluate the explanation methods. All code, models, and results are publicly available.

Table of Contents

List of Tables and Figures

Page

List of Symbols

Acronyms

AI	Artificial Intelligence
ADAM	Adaptive Moment Estimation
AUC-ROC	Area Under Curve – Receiver Operating Characteristic
CAD	Computer Aided Diagnostics
CNN	Convolutional Neural Network
DL	Deep Learning
GDPR	General Data Protection Regulation
Grad-CAM	Gradient Class Activation Mapping
ISIC	International Skin Imaging Collaboration
LIME	Local Interpretable Model-Agnostic Explanations
MCC	Matthews Correlation Coefficient
ML	Machine Learning
SDG	Sustainable Development Goal
SHAP	Shapley Additive Values
UKCA	United Kingdom Conformity Assessed
UN	United Nations
US FDA	United States Food and Drug Administration
UV	Ultraviolet
xAI	Explainable Artificial Intelligence

Acknowledgement

I would like to thank my family for their support. I would also like to thank Dr James Moore, Dr Andrea Marheim Storås, Dr Reiko Tanaka and Dr Abhuyday Tiwari for sharing their professional/academic expertise in this field of research.

Chapter 1: Introduction

1.1 Background

There is a lot of ongoing research on complex models that can ingest large amounts of data to support medical diagnosis decisions. Several highly accurate models trained in a variety of specific areas such brain disorders, breast cancer, skin cancer, gait issues, etc. can be found. Nonetheless, these models have found limited application in clinical settings. There are a few reasons for this, including a lack of trust from the humans in the loop and issues related to the generalization abilities of the models. To solve these problems, there has been research that implements explainable AI techniques to aid in improving trustworthiness. One of the issues with this research has been the gap in quantitative measurements of the effectiveness of these techniques.

1.2 Problem Statement

A skin lesion is abnormal skin growth. Some of these lesions can be cancerous, and their early detection is crucial for patient survival. Malignant melanomas, caused by UV exposure and commonly seen in areas with high sun exposure, can be deadly if not found early. The 5-year survival rate drops to 18% from 98% if the cancer spreads to other organs. Skin cancer rates around the globe are also on the rise. For the diagnosis of lesions, highly trained dermatologists are needed to look at the images. So, deploying accurate computer aided diagnosis tools can aid with the important issue of catching the tumor early on and reducing the clinical workload. Such models can also help reduce the incidence of biopsy tests, which are invasive and time consuming. Given these obvious benefits, there has been a rise of machine learning in medical diagnostics that has led to significant advancements in disease detection, particularly for skin cancer. For instance, leveraging the HAM10000 dataset, researchers have developed a number of classification models, including traditional machine learning techniques like Support Vector Machines, and Decision Trees as well as deep learning architectures like DenseNet, MobileNet, and ResNet (Tschandl et al., 2018). These models are slowly but surely moving to the implementation phase, with a large number of methods and devices approved by standards agencies, and AI/ML/DL basics being introduced in the education of clinicians.

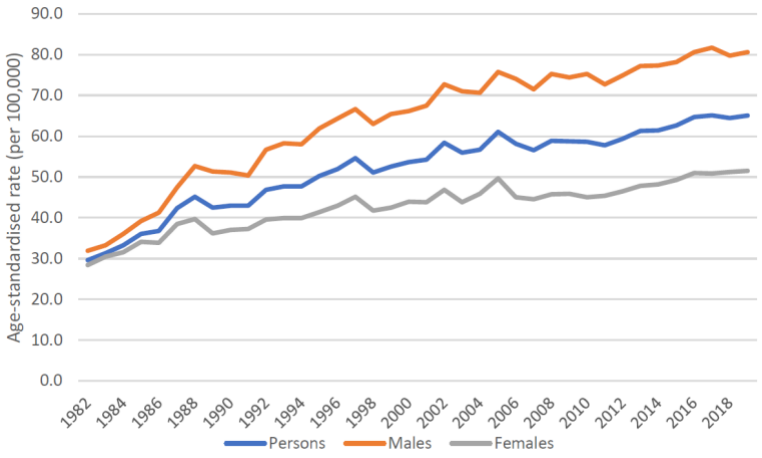


Figure 1: Rising Skin Cancer Incidence Rates in Australia

But many concerns persist about the reliability, interpretability, and transparency of these models. For example, a medical expert cannot easily reason through the output of these models. While many models have been shown to successfully classify skin lesions, matching and even outperforming human experts, the challenge to understand the reasoning behind these decisions remains (Haenssle et al 2018, Brinker et al 2019). Despite achieving a high accuracy, they are “black boxes”, with no human understandable insight available into their reasoning. This opacity is problematic in medical contexts where understanding the reasoning behind a diagnosis is crucial for clinical acceptance and patient trust (Amann et al., 2020). This can be incrementally problematic in scenarios where the models provide an outcome that differs from the medical expert. As noted in Holzinger et al (2017), new policies such as the changes to the GDPR are

now also making retrace-ability of decisions to be a requirement. Explainability will therefore be key in enabling deployment to clinics.

The emergence of Explainable AI (xAI) techniques like SHAP, LIME, and Integrated Gradients promises to demystify these models by offering post hoc explanations (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017). However, there is a significant gap in the literature regarding the quantification of the effectiveness of these xAI methods. Current research predominantly focuses on the implementation of xAI without critically assessing their explanatory quality, leading to a need for systematic evaluation frameworks. There have been many explainability studies that focus on showing features that influence the model's decisions for example. But there are barely any that focus on how well the visual technique explains the model and its outputs.

1.3 Model

For training to classify various cancerous and non-cancerous lesions, the proposed model uses an existing well-known dataset used in this field. Namely, the ISIC dataset (2018): HAM10000. This dataset consists of images of skin lesions along with labels describing whether or not the lesion is cancerous and which specific major category it belongs to if so. The models trained on this dataset are: Convolutional Neural Networks, DenseNet, MobileNet, Inception, etc. Once the models were trained, model agnostic explainability techniques like Integrated Gradients, Shapley Values and LIME were used to train explainers on top of the models. Metrics that tested for faithfulness and robustness were then used to quantitatively compare the explainers and therefore improve them through an improved choice of hyperparameters.

1.4 Solution Approach, Outcomes, and Contribution

To address the challenge of enhancing the trustworthiness of AI-driven skin cancer diagnostics, this research employed a multi-faceted approach using the HAM10000 dataset. Initially, various classification models were developed, including simple convolutional neural networks and other, more complex deep learning architectures. Each model underwent rigorous training with data augmentation and oversampling to improve accuracy, precision and recall. Hyperparameter tuning ensured optimal model performance. Subsequently, Explainable AI (xAI) methods, including SHAP, LIME, and Integrated Gradients, were applied to generate post hoc explanations for model predictions. The novel contribution of this research lies in the quantitative evaluation of these visual explanation methods using a variety of different metrics focused on faithfulness, robustness, localization, etc. This comprehensive framework advanced the systematically by systematically assessing the explanatory power of xAI techniques under a number of metrics belonging to faithfulness, robustness and complexity. The lack of a diverse dataset and the complexity and computational intensity of these models represent significant limitations that require further work.

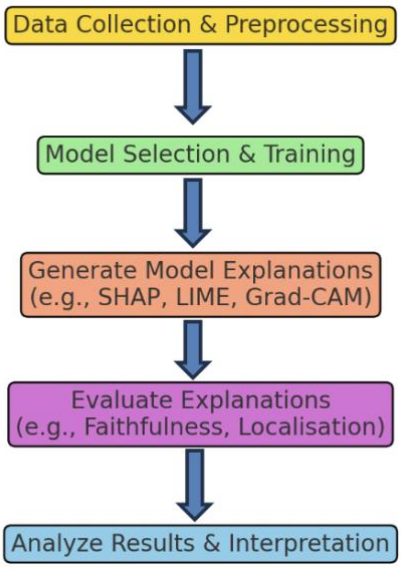


Figure 2: Solution Approach

The success of the proposed methods was evaluated through a well-designed experimental setup. The HAM10000 dataset was split into training, validation, and test subsets to ensure that the model could be validated and improved while training but also go through a final test with completely unseen data from a different source. Data augmentation and oversampling techniques were employed to address class imbalances and enhance generalizability. Custom code was developed to implement various models and xAI methods, followed by extensive hyperparameter tuning. Evaluation of the models involved common metrics in medical diagnostics such as accuracy, AUC-ROC, F1 score, recall, precision, and MCC. Additionally, novel explainability metrics such as faithfulness correlation and sensitivity were used to assess the explanations provided by the trained explainers. This dual approach of evaluating both model performance and explainability ensures a holistic assessment, which is a crucial part of the deployment of AI in medical diagnostics.

The research yielded significant and impactful results. The developed classification models achieved high accuracies. More importantly, the quantitative evaluation of xAI methods demonstrated how changing the parameters while training explanation techniques showed a clear correlation with changes in their explainability metrics. This directly enables the development of better explainers of the AI models. By testing some of these explainability metrics, this study helps work towards a robust framework for evaluating the explainability of AI models in medical diagnostics. Inductive bias introduced due to the model choices and the data used are limiting factors for the outcomes of this research. But the potential impact of exploring the methods above will be an increase in clinical deployment of these models, owing to an improvement in trust, interpretability, and reliability. Additionally, successful results in skin cancer diagnostics can also be replicated in other forms of image based diagnostics.

1.5 Outline

This project proposal is organized as follows. Chapter 2 reviews the literature on various types of models and key ideas, explainable AI methods and evaluation metrics for explainability, all within the context of applications in medicine. Chapter 3 further defines the problem. Chapter 4 describes the solution approach and methodology. Chapter 5 goes through the various legal, professional and ethical issues related to this research. Chapter 6 presents the results and finally, Chapter 7 provides the conclusion.

Chapter 2: Literature Review

2.1 Introduction

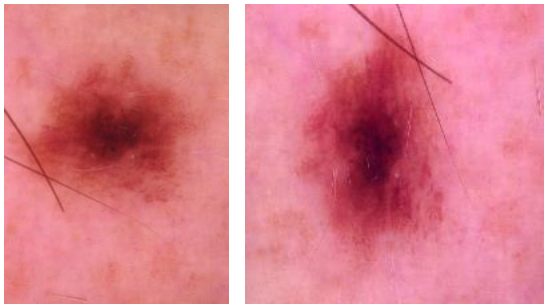
This literature review critically examines various models and techniques that address reliability and transparency concerns, focusing on fundamental concepts important to this study, along with previous applications in healthcare. The first sub-section focuses on the models (convolutional neural networks and transfer learnt DenseNet, MobileNet, etc.), including all the parameters important for the training. Next, is a look at the basics of Explainable AI and then its application in healthcare. Finally, the focus is on Evaluation of Explainability, including previous applications in healthcare.

2.2 Models

2.2.1 Model Parameters and Data

i. Learning Rate and Adjustments: Learning rate is a model hyperparameter that represents the step size per iteration as the optimizing function moves towards minimum loss. An appropriate learning rate is important for the model to actually converge. If it's too high, the model may converge quickly but to a suboptimal solution, but if it's too low, we can run into a long training process, face computing and resource constraints while still potentially getting stuck in a local minima. Techniques such as learning rate scheduling, where the learning rate is adjusted during training, can help in finding a balance. Reduce on plateau is a learning rate scheduling technique where the rate is reduced if the model's performance metric (such as validation loss) does not improve for over a specified timeframe/epochs. This approach helps the model converge more effectively by allowing finer adjustments when the training progress stalls (Bengio, 2012). While some research does suggest that cyclically varying learning rates between a minima and a maxima can provide accuracy and efficiency benefits (Smith, 2017), for the purpose of this study, a more conventional approach was used. Early stopping is another regularization technique used for avoiding overfitting by halting training when performance degrades. This technique is particularly beneficial in medical diagnostics, where overfitting can lead to models that perform very well in training but poorly on unseen data, thereby reducing their clinical utility.

ii. Data Augmentation and Sampling: Image augmentation is a technique where the training dataset is artificially expanded by applying random transformations such as rotations, shifts, translations, flips and crops to the existing image. This process helps in improving the model's robustness and generalizability by exposing it to a variety of scenarios during training (Perez and Wang, 2017). In medical diagnostics, with a scarcity of large annotated datasets, data augmentation is invaluable for enhancing model performance and preventing overfitting. Oversampling addresses class imbalance by duplicating instances of the minority class to ensure that the model receives enough representation from the minority classes during training. For instance, Synthetic Minority Over-sampling Technique (SMOTE) create synthetic examples rather than simple duplication, further enhancing model robustness (Chawla et al., 2002). In this study, a random oversampling technique was employed.



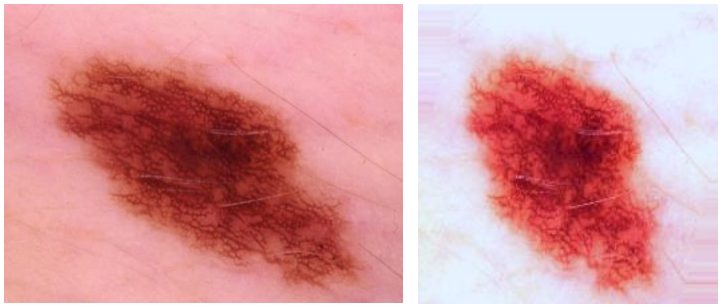


Figure 3: Some Sample Input and Augmented Images

iii. Loss and Optimizer Function: The choice of optimizer significantly impacts training dynamics and eventual performance of models. Stochastic Gradient Descent (SGD) has been a traditional choice due to its simplicity and efficiency but it can be slower to converge and more sensitive to the choice of learning rate. Adam (Adaptive Moment Estimation) is an optimizer computes adaptive learning rates for each parameter by considering the first and second moments of the gradients (Kingma & Ba, 2014). Adam's adaptive nature makes it more robust to the choice of hyperparameters and well-suited for handling complex, high-dimensional datasets common in medical diagnostics. Categorical cross-entropy loss is a loss function commonly used for classification tasks where the goal is to assign inputs to one of several categories. This loss function is particularly suitable for medical diagnostics, where models often need to differentiate between multiple disease states. Categorical cross entropy loss was used alongside an Adam optimizer for the purposes of this study.

iv. Epochs, Batch Size: The number of epochs determines how many times the model sees the dataset during training. Batch size refers to the number of training samples (images) used in one iteration. Selecting the right batch size and number of epochs is essential for model convergence. Too few epochs might lead to underfitting and too many can cause overfitting. Batch size affects the stability of the gradient descent process; smaller batches can provide a regularizing effect and reduce overfitting, but may lead to noisier gradient estimates (Goodfellow et al., 2016).

v. Model Evaluation Metrics: Accuracy, the proportion of correct results to all results, is a fundamental metric for evaluating model performance. But accuracy alone can be misleading, especially in datasets with class imbalance. Precision, the ratio of true positives to all positives (true & false), and recall/sensitivity (the ratio of true positives to the sum of false negatives and true positives) are essential metrics that provide deeper insights into a model's performance on imbalanced datasets. Precision is particularly important in medical diagnostics to ensure that a high proportion of identified positive cases are true positives, thereby minimizing false positives. For instance, Esteva et al. (2017) highlighted the importance of precision in their deep learning model for skin cancer classification to reduce unnecessary biopsies. Recall is crucial for identifying as many true positive cases as possible, thereby reducing false negatives. In applications like cancer diagnosis, high recall ensures that most cases of the disease are detected, which is vital for timely treatment.

F1 score, the harmonic mean of recall and precision, provides a measure that is more balanced, even when there is an uneven class distribution. The F1 score is a preferred metric in situations where the cost of false positives and false negatives needs to be equally minimized, such as medical diagnostics where both false positives and false negatives can carry significant consequences. MCC is another robust metric that provides a comprehensive evaluation of a model's performance. Unlike accuracy, MCC is high only if the classifier performs well across all confusion matrix categories, making it especially useful for imbalanced datasets. In medical diagnostics, MCC can provide a more accurate reflection of a model's diagnostic ability (Chicco & Jurman, 2020).

The AUC-ROC evaluates the trade-off between sensitivity (recall) and specificity (true negative rate) across different thresholds. AUC-ROC is particularly valued for its ability to provide a single metric that summarizes the model's performance across all classification thresholds. A model with a high AUC-ROC score indicates a good measure of separability, which is vital for reliable diagnostics. Multiple metrics need to be analyzed in tandem to get a good understanding of the capabilities of a model (Hicks et al, 2022).

Hyperparameter	Value
Learning Rate (Initial)	0.001
Optimizer Function	Adam
Loss Function	Categorical Cross Entropy
Loss Metric	Accuracy, F1 Score
Batch Size	32
Epochs	100
Augmentation	Random Flip, Zoom, Shift, Shear
Sampling	Random Minority Oversampling (rate=300%)

Table 1: An example of hyperparameters and values used during the study

2.2.2 Convolutional Neural Networks

CNNs represent a class of deep NNs designed for image processing and feature extraction. They automatically learn hierarchical representations from pixel data. The key innovation lies in the application of convolutional layers, which enable the network to efficiently capture local patterns and spatial hierarchies within images. Convolution layers consist of learnable filters that slide over the input image, performing convolution operations. This process allows the network to detect patterns, edges, and textures at different scales and orientations. Pooling layers downsample convolutional layers, reducing computational complexity while preserving important features. The final layers of a CNN integrate high-level features extracted by previous layers. These layers map these integrated features to the output classes through weighted connections. Activation functions, such as Leaky ReLU (Rectified Linear Unit) for example, introduce non-linearities to help the network handle more complex features.

Hosny et al. (2018) demonstrated the application of CNNs in skin cancer classification, employing transfer learning to leverage pre-trained models on large datasets. Alfi et al. (2022) and Alkarakatly et al. (2020) utilized CNNs for the diagnosis of melanoma skin cancer. The studies employed an ensemble stacking of machine learning models while also emphasizing the importance of explainability in the diagnostic process. The CNNs played a pivotal role in learning discriminative features from skin lesion images, contributing to accurate and interpretable predictions.

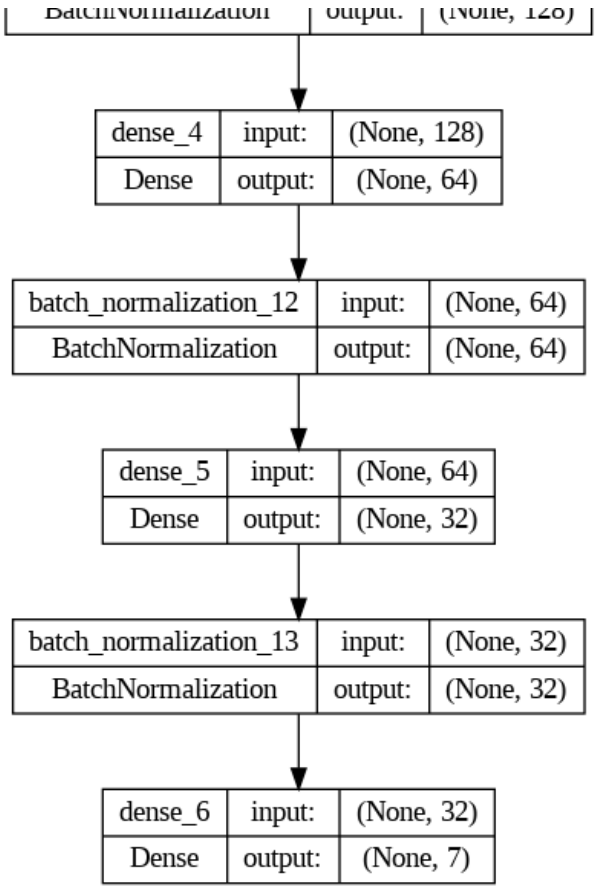


Figure 4: Last few layers of a CNN trained from scratch during the experimentation

2.2.2 Transfer Learning

Transfer learning uses a pre-trained model, trained on a large dataset such as CIFAR or ImageNet, and adapts it to a new, smaller dataset relevant to a specific task. This method leverages the pre-trained model's learned features, thereby reducing the need for extensive training and computational resources. This approach is beneficial in medical diagnostics, where annotated data is often scarce. Some network architectures used in transfer learning are described ahead.

1. DenseNet ensure maximum information flow by connecting each layer to every other layer in a feed-forward fashion. This architecture alleviates the vanishing gradient problem, encourages feature reuse, and improves model efficiency and accuracy (Huang et al., 2017). DenseNet has shown significant promise in medical image analysis, providing high accuracy with fewer parameters compared to traditional convolutional networks.

2. ResNet introduced the concept of residual learning, allowing networks to be substantially deeper by using identity shortcut connections that bypass one or more layers. This architecture addresses the degradation problem, where increasing network depth leads to higher training error (He et al., 2016). ResNet's ability to train very deep networks without performance degradation has made it a popular choice for various medical imaging tasks, including disease detection and classification (Brinker et al., 2018).

3. MobileNet is designed for efficiency, focusing on reducing the model size and computational requirements while maintaining high accuracy. It uses depthwise separable convolutions to reduce the number of parameters and operations (Howard et al., 2017). The Inception architecture aims to improve network performance by efficiently combining multiple convolutional filter sizes into a single layer output. This approach allows the network to capture features at various scales and increases its representational power (Szegedy et al., 2015).

Several different implementations of each of the models above are listed in Table 4. Fine-tuning is the process of further training a few or more previously ‘frozen’ layers of a pre-trained model. Weights are

slightly adjusted towards the new task, improving performance. Once some of the top layers of the model are unfrozen, it is jointly trained alongside the newly added classifier with a very low learning rate (Yosinski et al., 2014). In medical diagnostics, this allows for better characterisation of features specific to medical images, such as varying contrast, resolution, and anatomical differences. This process enhances the model's ability to generalize from the pre-trained domain (ImageNet) to the target domain (medical images), leading to better diagnostic accuracy.

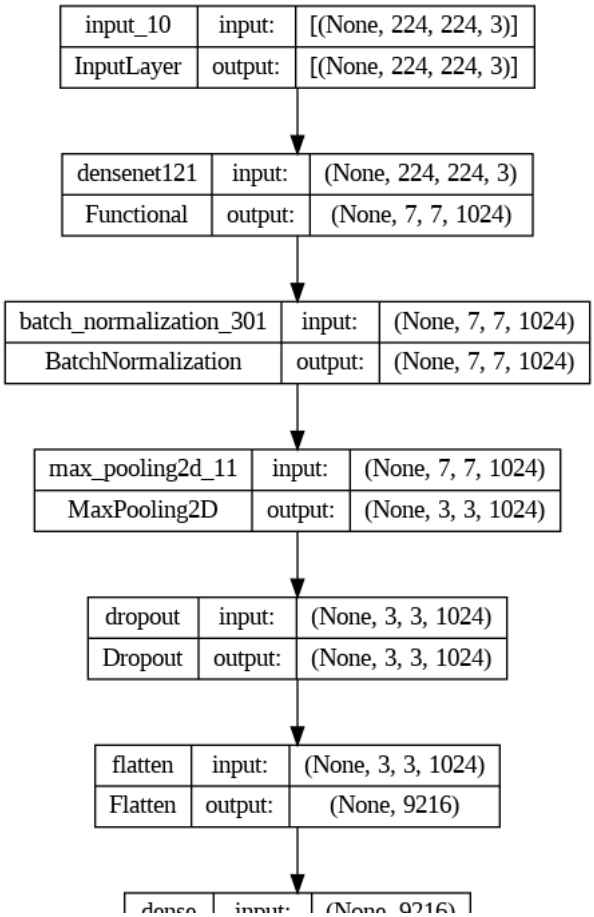


Figure 5: Some of the top layers of a transfer learnt DenseNet121 Model used in this study

2.3 Explainable AI

As artificial intelligence (AI) models, particularly deep learning models, become more pervasive in healthcare, there is an increasing need for explainable AI (xAI) to ensure that these models can be understood and trusted by clinicians. Explainability, interpretability, and trustworthiness are interrelated concepts that play crucial roles in the adoption of AI in clinical settings. This subsection examines various xAI techniques, distinguishing between explainability, interpretability, and trustworthiness, and explores both post hoc and ante hoc methods used to achieve explainability in AI models.

1. Explainability is the ability to provide understandable and transparent explanations for AI model outputs. This can involve detailing the inner dynamics of a model or generating insights that make its behavior comprehensible to humans.

2. Interpretability is often considered a subset of explainability focused specifically on making the model's predictions intelligible without necessarily exposing the model's internal mechanics.

3. Trustworthiness encompasses both the reliability of the model and the quality of its explanations. Trust in AI models is built through consistent performance, transparency, and robust explanations that enable users to verify and validate model predictions.

Post hoc methods generate explanations after the model has been trained. They do not alter the original model but instead provide insights into how the model makes decisions. These methods are often model-agnostic. Common post hoc techniques include:

1. Feature Importance: Techniques such as SHAP and LIME highlight which features influence the model's predictions the most. SHAP values are assigned to each feature, indicating its contribution to the model's prediction. Derived from cooperative game theory, SHAP values provide a fair distribution of credit among features. SHAP provides consistent, locally accurate explanations and can handle any machine learning model. It also offers a global perspective by aggregating local explanations (Lundberg & Lee, 2017). LIME generates interpretable models for individual predictions by perturbing input instances and observing the corresponding changes in predictions, identifying the most influential features in the process. It creates a local surrogate model to approximate the behavior of the original model around a specific prediction. This surrogate is typically a simple, interpretable model like a linear regressor or a decision tree which can be much easier to interpret (Ribeiro et al., 2016).

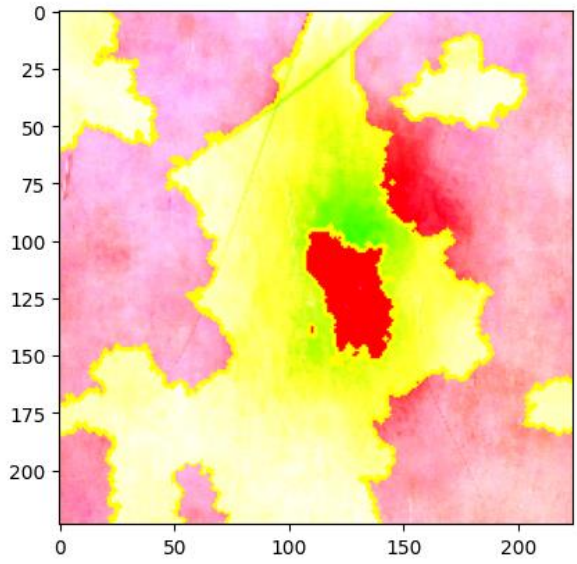


Figure 6: Example of a LIME Explanation. The green area signifies positive relationship with the label and the red signifies a negative on with the particular label under consideration. Axes are pixels, with the image of size 224 x 224

2. Saliency Maps: Used primarily in image classification, saliency maps visualize parts of the image which the model considers most important for its predictions. These maps are particularly useful in medical imaging applications.

3. Counterfactual Explanations: These provide alternative scenarios by showing how slight changes in input features could alter the model's prediction. This helps users understand the decision boundaries of the model.

4. Rule Extraction: Methods like decision trees or rule-based systems extract rules from complex models, making them more interpretable by translating model decisions into simple, human-readable rules.

5. Integrated Gradients: This method, particularly useful for neural networks, attributes the prediction of the network to its input features by integrating gradients of the model's output with respect to the input along the path from a baseline input to the actual input. This provides insights into which features are most important for the model's decisions (Sundararajan et al., 2017).

6. Grad-CAM: For CNNs, Grad-CAM generates visual explanations by highlighting parts of an image that are most important for a specific model output. It uses the gradients of the target concept (e.g., a

specific disease) flowing into the final convolutional layer to produce a coarse localization map of the important regions (Selvaraju et al., 2017).

7. Feature Visualization: Techniques like activation maximization help in understanding what kind of input patterns maximize the activation of certain neurons in neural networks. This method can provide insights into the hierarchical feature representations learned by the model.

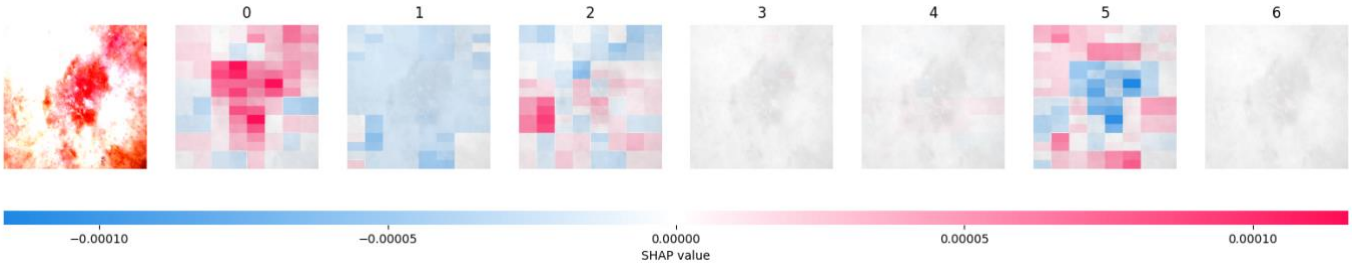


Figure 7: SHAP Example, here red denotes positive attribution towards the label and blue denotes negative. In this correctly classified image (label 0), you can see the features/regions that contributed to that output

Ante hoc methods integrate explainability into the model during the training phase. These models are inherently interpretable because they are designed with transparency in mind. Common ante hoc techniques include:

1. Interpretable Models: Models such as linear regression, decision trees, and logistic regression are inherently interpretable due to their simple structure. These models provide clear and straightforward explanations of their predictions.

2. Attention Mechanisms: highlight important parts of the input data, providing insight into which features are most relevant for the model's decision-making process.

3. Self-Explaining Models: These models, such as some types of neural network architectures, are designed to produce both predictions and explanations simultaneously. An example is the Neural Additive Model (NAM), which combines the flexibility of neural networks with the interpretability of additive models.

In this paper, the focus is primarily on Post Hoc Methods like SHAP, LIME and Integrated Gradients.

2.5 Applications in Healthcare – Examples of xAI

Wang (2023)'s survey on calibration in deep learning recognizes the significance of uncertainty quantification and xAI techniques in enhancing the explainability of models. Magesh et al. (2020) focuses on an explainable machine learning model for the early detection of Parkinson's disease using LIME. Singh et al. (2020) contributes to the understanding of interpretability in medical image analysis. Shahsavari et al. (2023) introduce an ensemble of deep models for skin lesion detection, emphasizing the importance of interpretable machine learning in dermatology. Pieter Van Molle et al. (2018) focus on visualizing CNNs to improve decision support for skin lesion classification. The study highlights the significance of interpretability in the context of dermatological applications. Cristiano Patrício et al. (2023) conduct a survey on Explainable Deep Learning Methods in Medical Image Classification. The review encompasses a wide range of xAI methodologies, offering a comprehensive perspective on the interpretability landscape in medical imaging.

Yaqoob et al. (2023) provide insights into the applications and techniques of machine learning in cancer classification. Xu et al. (2021) propose MANet, a two-stage deep learning method for the classification of

COVID-19 from chest X-ray images. Sobhan and Mondal (2021) focus on explainable machine learning to identify patient-specific biomarkers for lung cancer. Binder et al. (2021) explores morphological and molecular breast cancer profiling through explainable machine learning. Jiang et al. (2020) contribute to skin lesion segmentation based on a multi-scale attention convolutional neural network. Rezazadeh et al. (2022) explore explainable ensemble machine learning for breast cancer diagnosis based on ultrasound image texture features. The application of ensemble models and explainability metrics in this study offer valuable insights into enhancing the interpretability of models in dermatological applications. Ladbury et al. (2022) explore model-agnostic explainable artificial intelligence frameworks in oncology.

2.6 Explainability Evaluation Metrics

While various xAI methods have been developed on top of complex models, the evaluation of these methods' effectiveness remains a significant challenge. The effectiveness of xAI methods is typically evaluated using a range of metrics that assess different aspects of explanation quality. This review covers various approaches to evaluating the explainability of AI models, including user/expert studies, qualitative metrics, and automated quantitative metrics such as faithfulness, localization, complexity, plausibility, and robustness.

User/Expert Studies

User and expert studies are crucial for evaluating the interpretability and usability of AI explanations. These studies often involve domain experts and end-users assessing the explanations provided by AI systems to determine their comprehensibility and usefulness in decision-making processes. The feedback obtained helps in refining the models to better align with human cognitive processes and domain-specific requirements. For example, Doshi-Velez and Kim (2017) emphasized the importance of human-grounded evaluations, where explanations are tested on real users to assess their utility in practical scenarios. Poursabzi-Sangdeh et al. (2018) conducted studies to understand how explanations influence expert and non-expert users' trust in model predictions and their decision-making processes.

Qualitative Metrics

Qualitative metrics involve subjective assessments of explanations. These metrics consider how well an explanation communicates the model's reasoning process to users, including clarity, coherence, and completeness. Miller (2019) discussed various principles of human interpretability and highlighted the need for explanations to be simple, coherent, and contextually relevant to be effective. Ehsan et al. (2019) proposed a framework for evaluating the narrative quality of explanations, focusing on how well the explanation tells a story that users can understand and relate to their domain knowledge.

Automated Quantitative Metrics

Automated quantitative metrics provide objective measures for evaluating explainability. These metrics are particularly useful for large-scale assessments and benchmarking different models.

1. Faithfulness: measures how accurately an explanation reflects the true reasoning process of the AI model. It ensures that the explanation is a truthful representation of the model’s decision-making process. For example, faithfulness correlation evaluates the correlation between importance assigned by the explainer and the actual impact of those features on the model's predictions (Bhatt et al., 2020).

2. Localization: assesses how well an explanation identifies the relevant parts of the input. Localisation metrics usually rely on comparing segmentation maps with explainer identified image regions for assessing Intersection over Union for instance.

3. Complexity: measures the cognitive load required for a human to understand the explanation. Lower complexity indicates a more interpretable explanation. Rudin (2019) argued for simpler models and explanations, emphasizing that less complex explanations are easier for humans to understand and verify.

4. Plausibility: assesses whether the explanation makes sense to human experts, even if it is not a perfect representation of the model's internal logic. Kulesza et al. (2015) proposed evaluating explanations based on how plausible they appear to domain experts, regardless of their faithfulness.

5. Robustness: evaluates the stability of explanations under small perturbations of the input. An explanation is robust if small changes in the input do not significantly alter the explanation. Alvarez-Melis and Jaakkola (2018) introduced methods to quantify the robustness of explanations, ensuring they remain consistent under similar conditions.

Despite these advancements, several gaps remain in the evaluation of xAI methods. One significant gap is the lack of standardized benchmarks and datasets specifically designed for evaluating explainability metrics. This makes it challenging to compare the effectiveness of different xAI methods consistently. Additionally, while faithfulness and robustness are well-studied, comprehensibility remains subjective and challenging to quantify without extensive user studies, which are often resource intensive. Many existing studies also only focus on controlled environments or synthetic datasets, avoiding the variabilities encountered in practical applications (Bhatt et al., 2020).

2.7 Applications in Healthcare - Examples of xAI Evaluation

Study	Methodology	Metrics
Makridis et al. (2023)	Proposed a unified multidimensional explainability metric for healthcare AI models	User trust, Model transparency
Hu et al. (2022)	Developed an explainable retrieval system for medical images	Insertion and deletion based similarity scoring metrics
Jin et al. (2022)	Evaluated xAI algorithms on brain imaging tasks	Faithfulness, Plausibility, Physician feedback
Jin et al. (2023)	Provided guidelines and an evaluation framework for clinical explainable AI in medical imaging	Comprehensive evaluation framework
Zou et al. (2023)	Introduced an ensemble image xAI algorithm for severe pneumonia and COVID-19 diagnosis	Fidelity, Completeness, Clinical relevance
Luis A. de Souza et al. (2021)	Used saliency maps and other xAI techniques in cancer diagnosis	User study, Saliency maps
Pandey et al. (2024)	Introduced quantifiable xAI methods for cardiac disease diagnosis	Fidelity, Completeness, Stability
Ghanvatkar et al. (2023)	Integrated social science methods for assessing AI explanations in clinical practice	Quantitative metrics (Accuracy, AUC), Qualitative assessments (clinician feedback)
Siddiqui et al. (2023)	Developed trust metrics for evaluating deep learning models in medical time series classification	User trust, Model fidelity

Table 2: Summary of Key studies in Explainability Evaluation in Healthcare

The evaluation of xAI methods involves several metrics that measure different aspects of interpretability and trustworthiness. Key metrics include fidelity, completeness, consistency, stability, and user trust. These metrics help in understanding how well the explanations reflect the model's actual behaviour, whether they account for all significant features, and how robust they are against small perturbations in input data. But there is little consensus in the research as to which metrics need to be used for clinical

adoption. Makridis et al (2023) propose a framework that combines multiple dimensions of explainability, providing a comprehensive assessment that includes user trust and model transparency. Hu et al. (2022) developed an explainable retrieval system for medical images using insertion and deletion based similarity scoring metrics to grade different saliency maps of test datasets in ISIC and COVID, demonstrating the potential of explainable retrieval systems to enhance clinical decision-making by providing interpretable image comparisons. Jin et al. (2022) evaluated various xAI algorithms on brain imaging tasks, using faithfulness and plausibility metrics in the process along with feedback from physicians.

Evaluation metrics including fidelity, completeness, and clinical relevance were used by Zou et al (2023) to assess the effectiveness of their ensemble xAI models. The results underscored the potential of using ensemble xAI methods to improve robustness and reliability. Luis A. de Souza et al. (2021) used saliency maps and other xAI techniques to explain model decisions in diagnosing cancer and scored the methods by comparing them against a user study of human expert segmentation results. Pandey et al. (2024) introduced quantifiable xAI methods specifically designed for cardiac disease diagnosis. Metrics used included fidelity, completeness, and stability. Ghanvatkar et al. (2023) integrated social science methods to gauge usability and interpretability of AI explanations in clinical practice. The evaluation included both quantitative metrics (e.g., accuracy, AUC) and qualitative assessments (e.g., clinician feedback). This approach promises to provide a comprehensive evaluation of xAI methods, ensuring they met the practical needs of clinicians. Siddiqui et al. (2023) developed trust metrics for evaluating the explainability of deep learning models in medical time series classification. The research used ensemble methods to enhance the interpretability and robustness of AI models, employing metrics like user trust and model fidelity to assess their effectiveness.

From the above, we can note that combining quantitative metrics with qualitative assessments is crucial for a comprehensive evaluation of xAI methods. Equally important is the ensembling of different explanation methods together, an idea employed by at least 2 of the above-mentioned studies. Despite the progress in xAI evaluation in healthcare, several challenges remain. One major issue which is clear from the variety of metrics used in the aforementioned research, is the lack of standardized evaluation frameworks that can be universally applied across different healthcare domains and applications. Additionally, there is a need for more diverse datasets to ensure that xAI methods are generalizable and applicable to various patient populations. The integration of xAI methods into clinical workflows can be further complicated by the fact that end users sometimes prefer other explanation methods than those that are highest ranked by quantitative metrics. Ensuring that healthcare professionals trust and understand AI explanations is critical for the successful adoption of xAI methods in practice and so, it is important to integrate end users in the selection of the explanation model.

Chapter 3: Problem Statement

3.1 Introduction

Skin cancer remains one of the most common and dangerous forms of cancer globally. Accurate, early diagnosis is crucial for proper treatment and improving patient outcomes. Despite the considerable promise of AI/ML in diagnosing diseases, widespread adoption in clinical settings remains limited due to concerns surrounding trustworthiness and explainability. The primary aim of this thesis is to enhance diagnostic accuracy and provide meaningful explanations for model predictions using various explainable AI (xAI) methods. By quantifying and comparing the effectiveness of these xAI methods, this research seeks to improve the interpretability and reliability of skin cancer diagnostic models.

3.2 Problem Definition

Drawing insights from studies like Wang (2023), Magesh et al. (2020), Singh et al. (2020), Shahsavari et al. (2023), and Pieter Van Molle et al. (2018), this research underscores the critical role of explainable AI (xAI) methodologies. These methodologies contribute not only to the understanding of model decisions but also to the establishment of trust between healthcare professionals and machine learning systems. The project aligns with the vision outlined by Ladbury et al. (2022), exploring model-agnostic xAI frameworks in oncology, with metrics such as fidelity, faithfulness, and robustness offering quantitative insights into the reliability of skin cancer diagnostic models. The base classification problem in this study uses dermoscopic images from the HAM10000 dataset. The dataset comprises images classified into seven categories: Actinic keratoses (akiec), Basal cell carcinoma (bcc), Benign keratosis-like lesions (bkl), Dermatofibroma (df), Melanocytic nevi (nv), Melanoma (mel), Vascular lesions (vasc).

The task is a multi-class classification problem defined as:

- Input: Dermoscopic images
$$X = \{x_1, x_2, \dots, x_n\}$$
- Output: Corresponding labels
$$Y = \{y_1, y, \dots, y_n\}$$

where
$$y_i \in \{\text{akiec}, \text{bcc}, \text{bkl}, \text{df}, \text{nv}, \text{mel}, \text{vasc}\}$$

The primary hypothesis of this study is that by quantifying the effectiveness of xAI methods under different hyperparameters and comparing them, we can determine which methods provide the most reliable and interpretable explanations for skin cancer diagnosis models. This approach aims to fill the existing gap in evaluating the quality of visual explanations provided by xAI methods in medical diagnostics.

3.3 Model

This section presents the various models used, detailing their elements, decision variables, objectives, and constraints.

The following Loss and Optimizer Function equations are applicable to all models mentioned below:

Categorical Cross Entropy Loss:

$$L(y, \hat{y}) = -\sum(y_i \log(\hat{y}_i))$$

where y is the true label, $\hat{y}$ is the predicted probability, summed over the number of classes.

SGD Optimizer:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

where θ_t represents the parameters at step t, η is the learning rate, and ∇ is the gradient of the loss function J with respect to the parameters θ. Adam was used during training, but SGD equation shown first for simplification. Adam combines Adaptive Gradient Descent with Root Mean Squared Propagation. The steps for Adam are as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where α is the learning rate, β_1 and β_2 are the exponential decays for the first and second moments respectively, m_t is the first moment (mean of gradients), t is the time step, ν_t is the second moment (variance of gradients), θ_t are the model parameters at time t and ε is a small constant to prevent division by zero.

Models:

1. Convolutional Neural Networks (CNNs): Architecture: Simple CNNs designed from scratch including convolutional, max pooling, batch normalisation, dropout and dense layers. Decision Variables: Number of layers, filter sizes, and activation functions. Objectives: Minimize classification error and maximize diagnostic accuracy. Constraints: Computational resources and potential overfitting.
2. Transfer Learning Models (DenseNet, ResNet, and MobileNet): Architecture: Pre-trained models with added top layers and fine-tuned on the HAM10000 dataset. Decision Variables: Layers to freeze/unfreeze, learning rates, and batch sizes. Objectives: Utilize pre-trained knowledge to enhance performance on a smaller dataset. Constraints: Risk of overfitting and computational efficiency.

Model Equations and Explanation:

1. CNNs:

$$y_{pred} = softmax\left(W_f \cdot ReLU\left(W_{c_2} \cdot ReLU\left(W_{c_1} \cdot X + b_{c_1}\right) + b_{c_2}\right) + b_{c_f}\right)$$

While this is an oversimplification of a more complex CNN, the input image X is passed through two convolutional layers with ReLU activation, followed by a fully connected layer with SoftMax to output class probabilities. The weights and biases are what the classifier learns.

2. Transfer Learning:

$$y_{pred} = softmax\left(W_f \cdot ReLU\left(W_{c_2} \cdot TransferLearnt(DensetNet(X) + b_{c_2}\right) + b_{c_f}\right)$$

The input image X is processed through a pre-trained transfer learning model, followed by a fully connected layer to output class probabilities. Practically, a lot more top layers would be added onto the

model pre-trained on ImageNet weights and then some of the models layers would also be unfrozen for further fine tuning for improved accuracy.

Post Hoc Explainable AI Methods:

1. SHAP (SHapley Additive exPlanations): Quantifies the contribution of each feature to the final prediction, providing a global view of model behaviour

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(v(S \cup \{i\}) - v(S))$$

   where S is a subset of features, N is the set of all features and v(S) is the prediction when only features in S are present.

2. LIME (Local Interpretable Model-agnostic Explanations): Generates locally faithful explanations by approximating the model locally with an interpretable one

$$\xi = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

   where L is the loss function, f is the black-box model, g is the interpretable model, π is a proximity measure to the instance x, and Ω(g) is a complexity measure.

3. Integrated Gradients: Computes the average gradients of the output with respect to input features, providing insights into feature importance

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

   where x is the input, x′ is the baseline, and F is the model. The baseline can be a zero vector, black/white image, a blurred version of the input image or a random/uniform distribution. In this study, a random baseline was used (Fong et al, 2017).

Evaluation of Explainability

Next, we can look at some of the quantitative metrics commonly used for evaluating the explainability methods:

- Faithfulness is the correlation between importance score & actual impact of features

$$Faithfulness = Correleation(E_i, \Delta\, \hat{y})$$

where E_i is the explanation score for feature i and $\Delta\, \hat{y}$ is the change in the model prediction when feature i is perturbed or removed.

- Sensitivity evaluates how explanations change as the input changes, ensuring consistent identification of important features

$$Sensitivity = \frac{1}{n}\sum_{i=1}^{n} E(x_i) - E(x_i")$$

where $E(x_i)$ is the explanation for input x_i and $E(x'_i)$ is the explanation for a slightly perturbed input x'_i

- Infidelity is the difference between the explanation and the actual impact of feature perturbations

$$Infidelity = E_{x,\delta}[\ \ \delta \cdot E(x) - (f(x) - f(x-\delta)\ )\ )^2]$$

where δ is a perturbation, E(x) is the explanation for input x and f(x) is the model's prediction for input x.

- Monotonicity assesses whether increasing the importance score of a feature consistently leads to an increase in the model's prediction.

$$Monotonicity = \sum_{i=1}^{n-1} [\![[E(x_i) < E(x_{i+1})]\!] \Rightarrow f(x_i) < f(x_{i+1})]$$

where $E(x_i)$ is the explanation score, and $f(x_i)$ is the model prediction for input x_i

- Sparseness evaluates the simplicity of an explanation by measuring the proportion of features that are assigned a non-zero importance score. The idea is that a sparser explanation based on fewer features would be easier for humans to interpret.

$$S = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(S(x_i) > 0\right)$$

where $S(x_i)$ is the important score for feature x_i and n is the number of features

- Relative Stability measure consistency with respect to input perturbations. Input stability assesses consistency of feature importance scores while output stability measures consistency of the output

$$ROS = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{y}(x) - \hat{y}(x + \delta x_i)|}{\|\delta x_i\|}$$

where n is the number of perturbations δx is the magnitude of the perturbation, y is the model output. For the relative input stability case, y can be replaced with feature importance score S.

- IROF (Input Reduction Output Fidelity) measures how much of the input data can be reduced (removed or masked) while maintaining the model's original output. A lower IROF score indicates that fewer features are required to maintain the same output, suggesting a higher fidelity of the explanation.

$$IROF = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\hat{y}(x) = \hat{y}(x_{\setminus S_i})\right)$$

where n is the number of samples, x is input with features S removed and y is the output of the model.

Chapter 4: Methodology

4.1 Introduction

This section provides a comprehensive overview of the methodology used to address the problem of developing and evaluating explainable models for the diagnosis of skin cancer using the HAM10000 dataset. The methodology involves data preprocessing, classification model development, training, evaluation, and the application of explainable AI (xAI) methods to quantify and compare model interpretability.

4.2 Methodology

4.2.1 Data Collection and Preprocessing

The HAM10000 dataset comprises 10,015 dermoscopic images, categorized into seven classes: actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions (Tschandl et al., 2018). This dataset was chosen due to its class diversity, clinical relevance and the amount of previous classification research available on it for reference and comparison. The dataset was divided into training and validation sets using an 75-25 split, with an additional 1500 test images used to finally test the generalizability of the classifiers. Stratified sampling was employed to maintain class distribution across all sets. To enhance model robustness, image augmentation techniques such as channel and brightness shifts, shears, random rotations, flips and zooms were applied to the training set (Perez & Wang, 2017). Oversampling was used to reduce class imbalance and rates were based on previous research suggesting a max 300% increase in the smaller classes. Image pixel values were normalized to improve model convergence rates.

Preprocessing Pseudo-code:

$$
\begin{aligned}
&\#pseudo-code\ for\ data\ split\ into\ training\ and\ validation\\
&\text{def } data\_split(input = X, label = y):\\
&\quad X_{train},\ y_{train},\ X_{test},\ y_{test} = train\_val\_split(split\_rate = 0.25, input = X, label = y)\\
&\quad return\ X_{trian}, y_{train}, X_{val}, y_{val}
\end{aligned}
$$

$$
\begin{aligned}
&\#pseudo-code\ for\ oversampling\ minor\ classes\\
&\text{def } data\_oversampler(input = X, label = y):\\
&\quad X_{oversampled},\ y_{oversampled} = random\_oversampler\ (oversampling\_rate = 3, input = X, label = y)\\
&\quad return\ X_{oversampled}, y_{oversampled}
\end{aligned}
$$

$$
\begin{aligned}
&\#pseudo-code\ for\ image\ augmentation\\
&\text{def } augment\_generator(raw\_images):\\
&\quad augmented\_images = []\\
&\quad for\ img\ in\ raw\_images:\\
&\qquad img\ = resize(img)\\
&\qquad img\ = normalise(img)\\
&\qquad img\ = random\_channel\_shift(img)\\
&\qquad img\ = random\_flip(img)\\
&\qquad img\ = random\_zoom(img)\\
&\qquad img\ = random\_shear(img)\\
&\qquad img\ = random\_brightness(img)\\
&\qquad augmented\_images.append(img)\\
&\quad return\ augmented\_images
\end{aligned}
$$

4.2.2 Classification Model Development

CNNs were built from scratch with varying architectures, including different numbers of convolutional layers, kernel sizes, and activation functions. The architectures were optimized through hyperparameter tuning. Pre-trained models such as DenseNet, ResNet, and MobileNet were fine-tuned on the HAM10000 dataset. These models were chosen for their proven efficacy in image classification tasks.

Loss Function and Optimization: Categorical cross-entropy was chosen due to its suitability for multi-class classification problems. The Adam optimizer was used over Stochastic Gradient Descent (SGD) because of its adaptive learning rate capabilities, which typically result in faster convergence (Kingma & Ba, 2014). While SGD can perform better, it can be a lot slower to converge. Reduce on plateau was used to reduce the learning rate when the validation loss plateaued, allowing the models to converge more smoothly. Experiments were conducted with batch sizes of 8, 32, 64, and 128. The number of epochs varied between 50 and 200, depending on the convergence behaviour observed during training.

$$\#pseudo - code\ for\ classification\ models$$
$$\text{def } classifier\_training(input = X, label = y):$$
$$base\_model = DenseNet(weights = "ImageNet", without\_top\_layers)$$
$$base\_model.layers.train = False$$

$$x = base\_model.output$$
$$x = pooling\_layers(x)$$
$$x = dense\_layers(x)$$
$$x = dropout\_layers(x)$$
$$prediction = Dense(classes = 7, activation = softmax)$$

$$model.train(input = X, labels = y, output = prediction, validation = \{X_{val}, y_{val}\}, epochs,$$
$$batch\_size, optimiser = Adam(learning\_rate), loss = categorical\_cross\_entropy)$$

$$model.fine\_tune(base\_model.layers(top = 5))$$

$$return\ model$$

4.2.3 Explanation Model Development

Post hoc explainability methods such as Integrated Gradients, LIME and SHAP were applied to selected models to generate explanations for their predictions. These perturbations and saliency-based methods help in understanding the contribution of each pixel and feature to the decision of the model.

$$\#pseudo - code\ for\ explaination\ models$$
$$\text{def } explainer\_training(input = images, label = y, model = model):$$
$$shap\_explainer = ShapExplainer(model.predict, images, shap\_model\_parameters)$$
$$shap\_values = shap\_explainer(images)$$

$$lime\_explainer = LimeExplainer(model.predict, images, lime\_model\_parameters)$$
$$lime\_values = lime\_explainer(images)$$

$$integrated\_gradients\_explainer$$
$$= IntegratedGradientsExplainer\ (model.predict, images, int\_grad\_model\_parameters)$$
$$integrated\_gradients\_values = integrated\_gradients\_explainer(images)$$

$$return\ explainers, explanations$$

4.2.4 Model Evaluation

Text based explanations and their accompanying evaluation using methods such as METEOR (Metric for Evaluation of Translation with Explicit Ordering), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Understudy Evaluation) has been seen several times before (Patricio et al., 2023). This paper focuses instead on image based explanation methods such as Integrated Gradients, LIME and SHAP and evaluates their performance quantitatively.

Classification Models were evaluated using standard metrics such as accuracy, MCC, recall, precision, AUC and F1 score. These metrics comprehensively assess the performance of the model, particularly in the context of imbalanced datasets.

The effectiveness of xAI methods was quantified using metrics such as faithfulness correlation, sensitivity, stability, etc on different variations of the explainers for comparisons. These metrics help in assessing how well the explanations align with the classification model's actual resolution process.

$$\#pseudo - code\ for\ evaluation\ of\ explaination\ models$$
$$\text{def}\ explainer\_evaluation(explainer, model, explanations, ground\_truth):$$
$$metrics = \{\}$$

$$metrics.faithfulness = Faithfulness(model, explainer, explanations, ground_{truth}, runs = 10)$$
$$metrics.robustness = Robustness(model, explainer, explanations, ground_{truth}, runs = 10)$$
$$metrics.complexity = Complexity(model, explainer, explanations, ground_{truth}, runs = 10)$$
$$metrics.plausibility = Plausibility(model, explainer, explanations, ground_{truth}, runs = 10)$$
$$metrics.localization = Localization(model, explainer, explanations, ground_{truth}, runs = 10)$$

$$return\ metrics$$

Further details about the equations and theoretical foundations related to the implementations described above can be found in the literature review and problem statement sections.

4.2.5 Experimental Setup

Experiments were conducted on a Google Cloud Workbench with 1 NVIDIA L4 GPU along with 4 2-core vCPUs with 16 GB memory to accelerate the training and evaluation processes. The models were implemented using TensorFlow and Keras frameworks. Code was written from scratch as well as adapted from publicly available repositories as needed, with modifications made to suit the specific requirements of this project. Appropriate references to original authors were provided where necessary in the codebase. The code and trained models, along with documentation, were made available on GitHub (Sangwan 2024).

Chapter 5: Legal, Professional and Ethical Issues

6.1 Introduction

This chapter explores the professional and ethical considerations inherent in developing models for skin cancer diagnosis, emphasizing adherence to international regulatory standards and guidelines such as ISO standards for medical devices, CE marking, UKCA marking, and FDA certification. These regulations ensure that AI technologies are safe, trustworthy, and respect fundamental rights. The ultimate objective is to develop AI models that are not only technically robust but also compliant with these stringent standards, ensuring their safe and ethical deployment in clinical settings. *Please note that all compliance standards mentioned hereafter provide a reference for adherence, but in the course of building the model, I did not officially apply for or get certified through any of these standards.

6.2 PESTEL Analysis

Political: Regulatory frameworks like FDA guidelines, ISO standards and the EU AI Act provide a structured approach to developing AI-driven medical devices. These regulations mandate transparency, and accountability, ensuring that AI systems are safe and reliable. Political support for AI in healthcare can facilitate funding, regulatory approvals, and broader acceptance of these technologies.

Economic: AI-driven diagnostics can significantly reduce costs through automated, early, accurate diagnosis, thus reducing the need for more expensive treatments. The economic benefits include cost savings for healthcare providers and improved accessibility for patients. This model is in a very preliminary stage, but if we consider future development, it is worth noting that the costs of development, regulatory compliance, and implementation can be substantial. Ensuring compliance with ISO standards and obtaining certifications like CE marking and FDA approval can also add to these costs but are essential for market access and trust.

Standard/Certification	**Description**
ISO 13485:2016	Medical devices — Quality management systems — Requirements for regulatory purposes
ISO 14971:2019	Medical devices — Application of risk management to medical devices
ISO 14001:2015	Environmental management systems
General Data Protection Regulation (GDPR)	Data protection and privacy regulation in Europe
European AI Act	Regulation on AI within the EU
CE Marking	Conformity marking required for marketing medical AI systems in Europe
UKCA Marking	Similar to CE but specific to the UK Market
US FDA Certification	Approval of AI Systems in Healthcare in the US
ISO 9241-171:2008	Ergonomics of human-system interaction, Part 171: Guidance on software accessibility
ISIC Paper (Daneshjou et al 2021)	International Skin Imaging Collaboration Guidelines for Image based Dermatology AI

Table 3: List of pertinent standards/certifications for this study

Social: This project aims to democratize access to high-quality diagnostic tools, potentially bridging gaps in healthcare availability, especially in underserved areas. However, the current data lacks sufficient diversity, limiting the model's helpfulness across different demographic groups. ISIC released a more robust dataset in July 2024, containing 400,000 images from a diverse population sample, that should be used in future training of any such model. Future compliance with standards like ISO 13485 could ensure

that these tools meet high safety and quality benchmarks, addressing social concerns about AI in healthcare.

Technological: The use of advanced AI techniques like CNNs and transfer learnt DenseNet models has represented a crucial step forward in medical diagnostics. Although these techniques were already being researched as early as 5 years ago, gap in deployment remains. Making sure the models are understandable, which is a primary goal of this study, will go a long way in increasing clinical adoption. Additionally, ensuring robust cybersecurity measures is important to protect sensitive patient data and maintain trust in AI systems. The dataset used for this study is publicly available and has already gone through rigorous measures before being released by the International Skin Imaging Collaboration. Testing and validation of AI systems are necessary to ensure their reliability and safety before being released for public use.

Environmental: The environmental impact of AI research, particularly the computational resources required, should not be overlooked. Although transfer learning does partly improve training efficiency as opposed to building models from scratch, efficiency was not a focus in this current study. Future iterations should prioritize researching more efficient algorithms and using data centres with renewable energy sources to mitigate the environmental footprint.

Legal: Compliance with data protection laws, such as GDPR, is critical to safeguarding patient privacy and data security. Legal considerations also include intellectual property rights and ensuring that AI models are used ethically and responsibly. Adhering to standards such as ISO 14971 for risk management in medical devices can ensure that legal and safety risks are appropriately managed. No personal data was collected directly for this study and the dataset used passes all compliance requirements. The model created will not be released for public use before passing the various certifications and standards mentioned above.

6.3 Ethical Considerations

AI models must prioritize patient safety and well-being, adhering to ethical principles that respect human life. Minimizing the environmental impact of computational resources is essential to ensure sustainability for future generations. Compliance with ISO 14001 for environmental management can help mitigate these impacts. Ensuring robust data protection measures is paramount in handling sensitive medical data. Compliance with GDPR safeguards patient information, maintaining confidentiality and trust. The dataset used was anonymized and secure storage practices were followed on Google Cloud Storage during the manipulation of the data.

6.4 Inclusive Engineering Outcomes

The project should consider the perspectives of all stakeholders, including patients, healthcare providers, and policymakers. I personally talked to many doctors and other researchers in this field from the UK, India and the US while working on the project to get a better understanding of the situation on-site, patient needs and doctor's requirements/concerns. In the future, I would consider further contact with additional stakeholders like patients and policymakers. Engaging with these groups ensures that the solutions are relevant, effective, and widely accepted.

Figure 8: UN SDGs pertinent to this project

The ISIC group provides a checklist for evaluation of image-based AI algorithm reports in dermatology (Daneshjou et al 2021). This study follows all the data, technique, technical assessment and application guidelines recommended. The project anticipates future trends in healthcare and technology, ensuring that the solutions remain relevant and adaptable. The current model was not trained on a diverse dataset, and fairness audits were not conducted due to a lack of better data and time constraints. However, future efforts will focus on preventing bias and discrimination by training on diverse datasets and conducting regular audits for fairness. Efforts should be made to ensure that diagnostic tools are accessible to all, regardless of socioeconomic status or geographic location. Compliance with ISO 9241-171 for accessibility requirements can ensure that the tools are usable by people with diverse abilities. This project aligns with several UN SDGs, including Good Health and Well-being (Goal 3), Industry, Innovation, and Infrastructure (Goal 9) and Reduced Inequalities (Goal 10). By improving access to diagnostic tools and reducing health disparities, the project contributes to these global goals.

6.5 Conclusion

By keeping in mind regulations in standards such as ISO standards, CE marking, and FDA certification, we could ensure that the development of AI-driven diagnostic tools is responsible and sustainable. The research supports the UN Sustainable Development Goals by promoting health equity and innovation. This holistic approach underscores the importance of integrating ethical considerations into engineering practices, ensuring that technological advancements benefit society as a whole.

Chapter 6: Results

6.1 Introduction

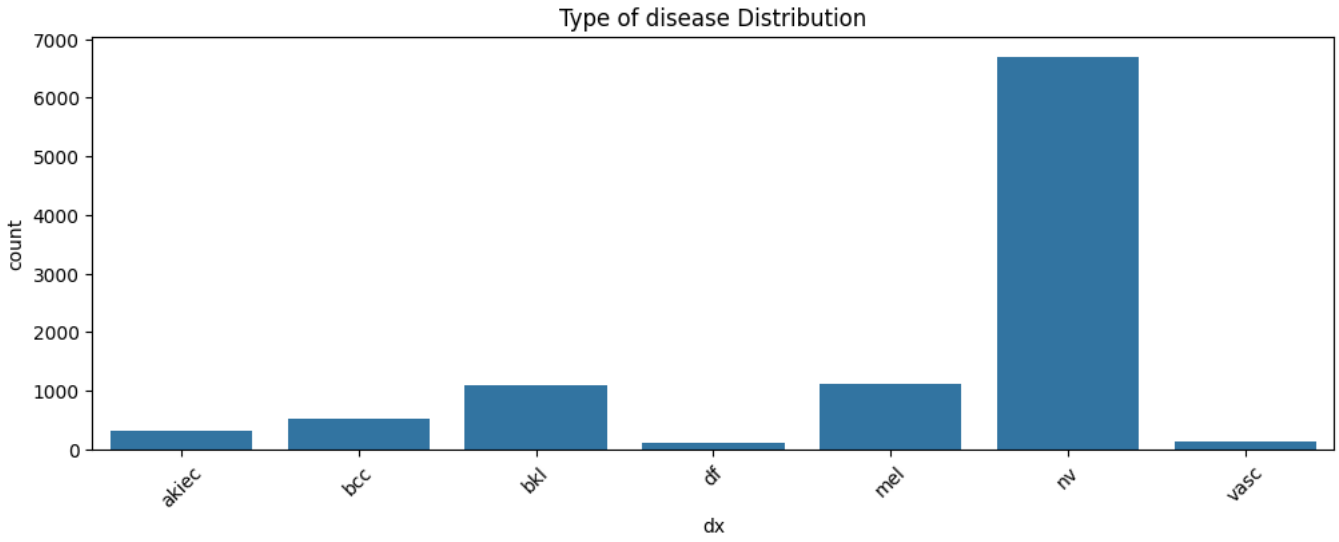
This chapter presents the results obtained from the experimental processes discussed in the previous chapters. The key findings are organized into sections focusing on data analysis, impact of oversampling and augmentation techniques, performance of the trained classification models, and evaluation of explanation methods. This analysis allows for a thorough understanding of the success and limitations of models and methods applied in this study.

6.2 Data Analysis and Oversampling

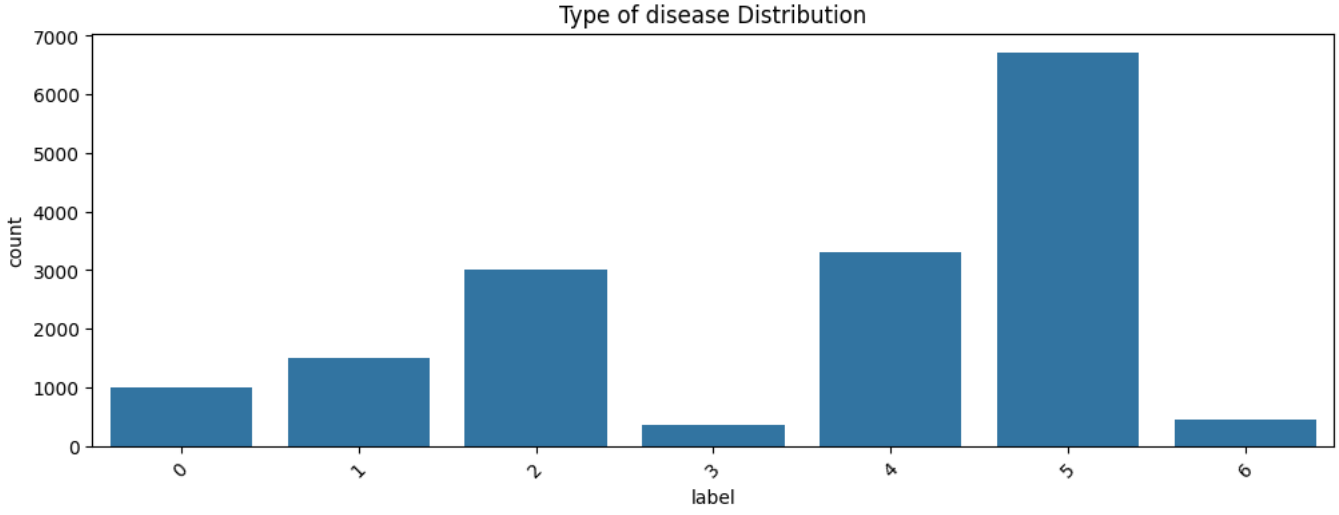
The dataset, HAM10000, initially exhibited significant class imbalance, with some classes being underrepresented. To mitigate this, oversampling techniques were employed, particularly focusing on randomly replicating instances from the minority classes.

Original Dataset: The original dataset distribution revealed a significant skew towards certain classes, potentially leading to biased model training. This imbalance posed a challenge in accurately predicting the underrepresented classes.

Oversampled Dataset: After applying random oversampling, the dataset was more balanced, ensuring that minority classes were adequately represented during the training process.



A) Original Dataset



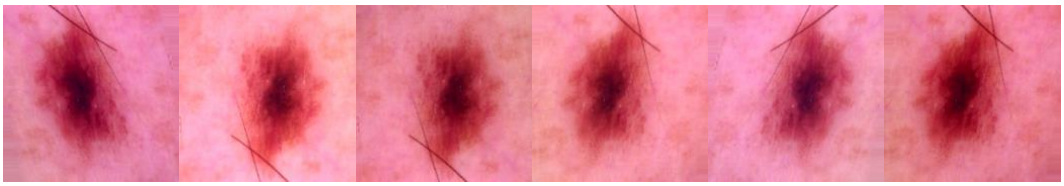
B) Oversampled Dataset

Figure 9: Comparing the distribution of the original vs the oversampled dataset

6.3 Image Augmentation

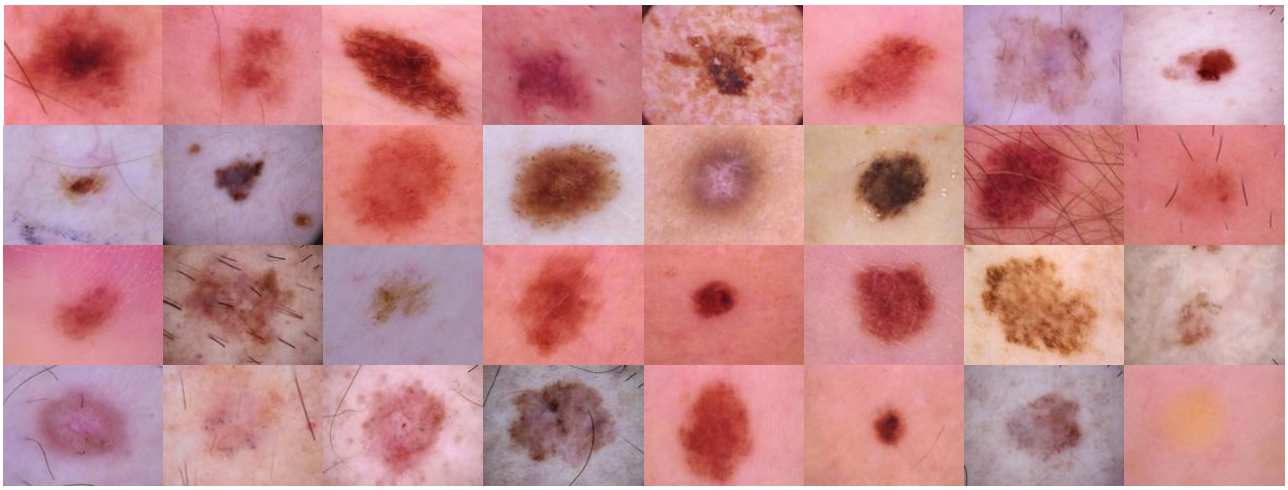
Augmentation involved generating multiple variations of input images through random transformations, such as rotations, translations, flips, and zooms. This approach increased the effective size of the training dataset while exposing the model to a wider variety of scenarios. A new random augmentation is applied to every image in each epoch. The two figures below showcase the augmentations. The first figure shows how the same image is augmented slightly differently in every epoch, thereby giving the model a bit more information. The second figure showcases how the augmentation is applied to a batch of images (batch size of 32 shown in the example).

a) Original Image

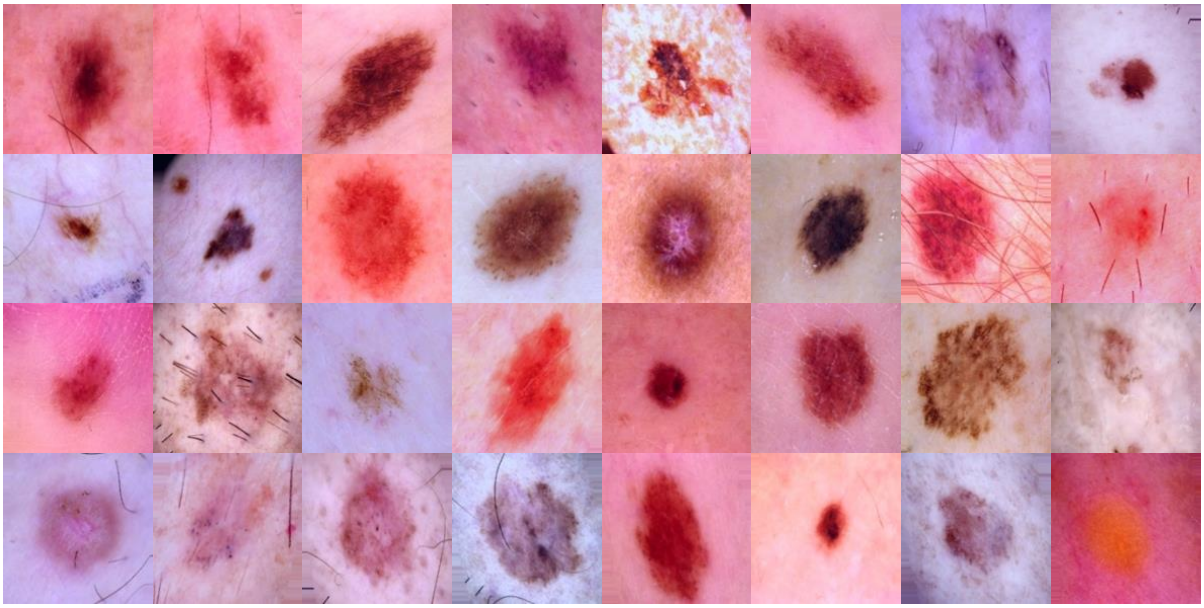


b) Augmentations of the Image

Figure 10: Example of different augmentations of the same image over different epochs. First image is the original followed by 6 different random augmentations over 6 epochs



a) Original Images



b) Augmented Images

Figure 11: Example of a batch of original images and their augmentations during a single epoch. New augmentations generated each epoch

6.4 Classification Models

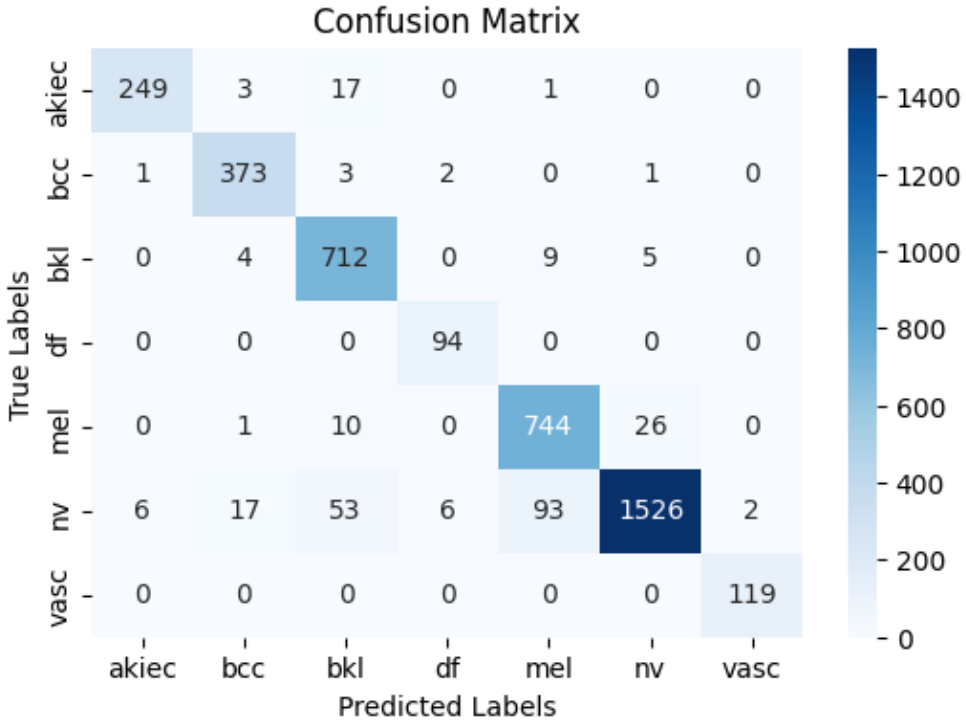
The study developed and tested several classification models, each trained and validated using the augmented and oversampled dataset. The models included a custom-built CNN as well as pre-trained models such as DenseNet and MobileNet, fine-tuned through transfer learning. The performance of these models was evaluated using a set of metrics, with the results shown below in comparison with other state-of-the-art research in this field. Please note that the results are from the validation set. Tested on a completely unseen set from a different source of images, the model achieved accuracy, precision and recall values of around 80%.

Model	Accuracy	Precision	Recall	F1 Score	AUC	MCC
CNN (from scratch)	87	87	87	87	98	81
MobileNet (transfer learning)	91	91	91	91	99	88
DenseNet (transfer learning)	93.28	93.62	93.28	93.3	99.64	92
Efficient-B4 (Huang et al. 2021)	85.8	91.91	96	93.91	-	-
ResNet50V2/EfficientNet-B0 (Bansal et al 2022)	88	86	89	87.48	-	-
Fully Convolutional Network DenseNet (Adegun and Viriri, 2020)	98.3	98	98.5	98	-	-

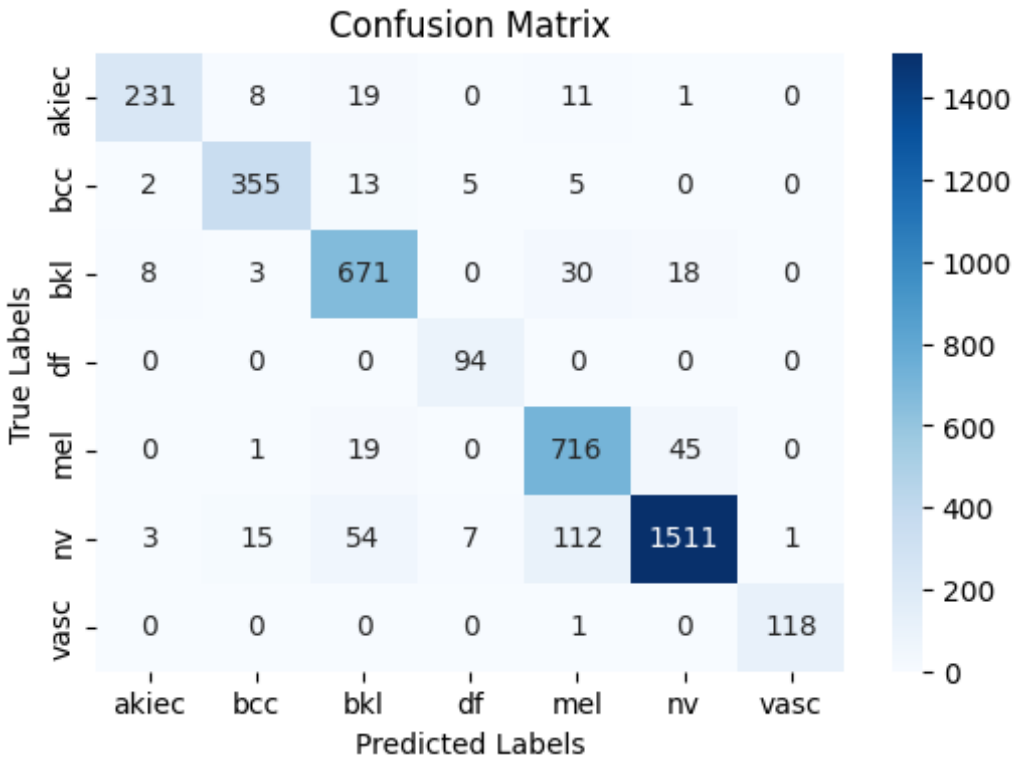
Generative Adversarial network (Qin et al 2020)	95.2	74.3	83.1	78.45	-	-
InceptionV3 (Rab Ratul et al)	90.95	89	89	89	-	-
Genetic Algorithm tuned CNN (Salih and Duffy, 2023)	95.96	96	95.86	96	-	-
VGG16 (Tabrizchi et al 2022)	87.05	-	85.23	92.21	92.31	-
DermoExpert (Hasan et al 2022)	-	85	86	-	97	-
Hybrid U-Net MobileNetV3 (Lilhore et al 2024)	98.03	97.32	94.07	97.03	97.09	-
ResNet101 Dynamic Multiscale CNN (Han et al 2024)	97.3	96.58	94.94	94.78	-	-
Ensemble Deep Learning (Hossain et al 2024)	95	95	95	95	-	-
Improved Recurrent Unit Networks and Orca Predation Algorithm (Zhang et al 2024)	98	95	95	95	-	-

Table 4: Comparing the trained classification models' validation metrics with state-of-the-art methods

The figures below show a more detailed view of the models' performance, showing some examples of the class wise metrics, confusion matrices and training history.



A) DenseNet



B) MobileNetV3

Figure 12: Confusion Matrices on Validation Data

Classes	**Precision**	**Recall**	**F1-Score**
Akiec	0.97	0.92	0.95
Bcc	0.94	0.98	0.96
Bkl	0.92	0.98	0.93
Df	0.88	1.00	0.96
Mel	0.98	0.95	0.91
Nv	0.98	0.90	0.94
vasc	0.98	1.00	0.99
accuracy	0.94		
Macro avg	0.94	0.96	0.95
weighted avg.	0.94	0.94	0.94

Table 5: Class-wise metrics example (DenseNet121)

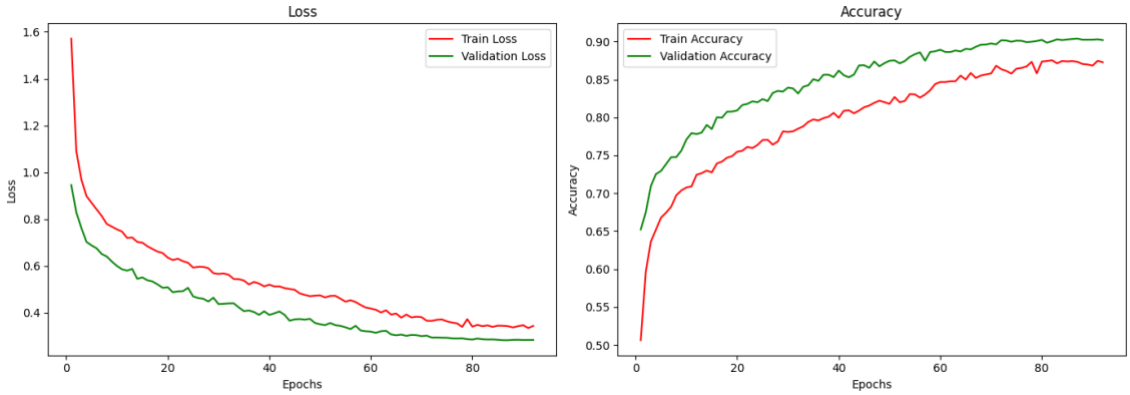
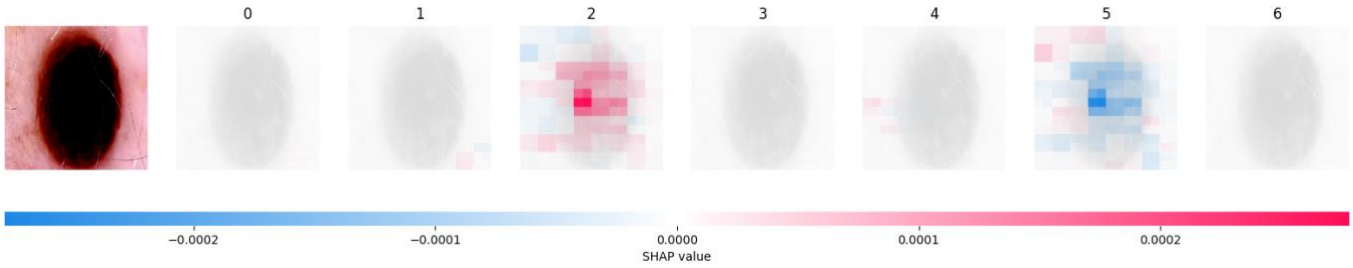


Figure 13: Training History Example (MobileNetV3)

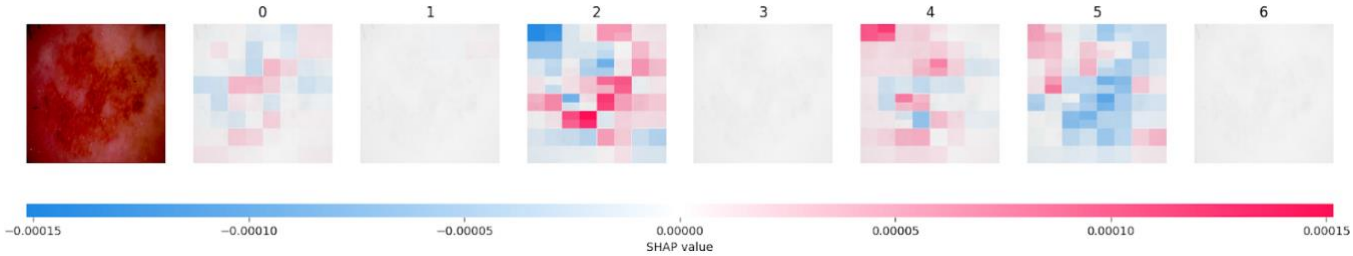
6.5 Explanation Models

Post hoc explanation methods, including Integrated Gradients, LIME and SHAP were applied to the predictions made by the classification models. These methods provided insights into which features contributed to the models' decisions. Some examples of this implementation are shown below.

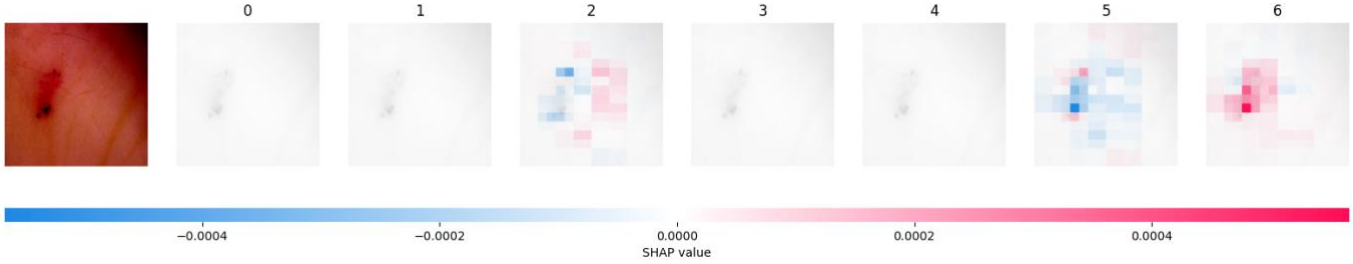
SHAP Examples



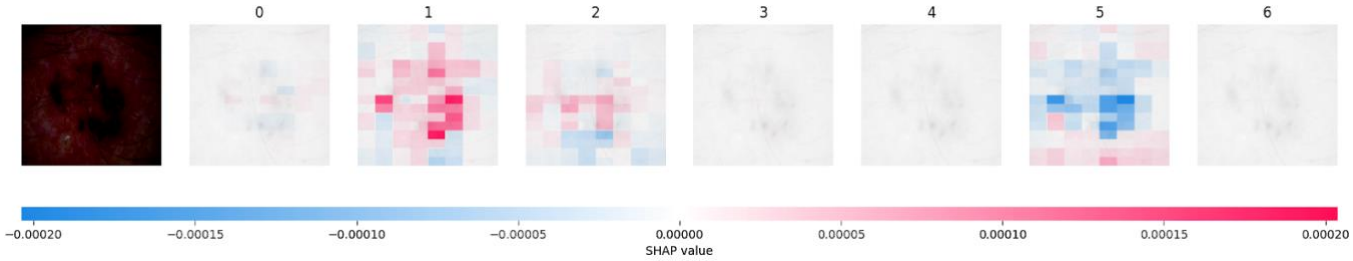
A) Predicted Probabilities: bkl/’2’: 9.8544139e-01, nv/“5”: 1.4558283e-02



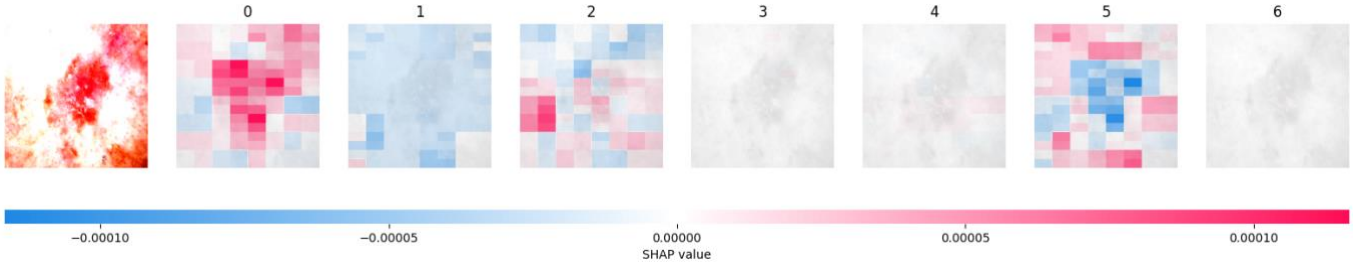
B) Predicted Probabilities: mel/”4”: 5.0333261e-01, bkl/“2”: 2.3027517e-01, nv/”5: 2.6623964e-01



C) Predicted Probabilities: vasc/”6”: 9.9994862e-01



D) Predicted Probabilities: bcc/“1”: 8.5888600e-01, nv/”5”: 1.2207900e-01



E) Predicted Possibilities: akiec/“0”: 9.8774838e-01

Figure 14: Some Examples of SHAP with the original image on the left and the accompanying map of SHAP values for each label sequentially following it. Red indicates positives markers for that label while blue indicates markers against that label.

LIME Examples

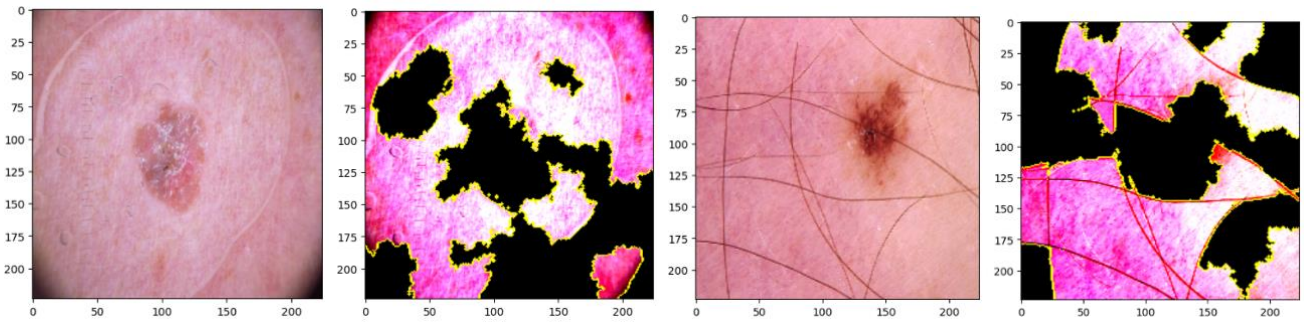
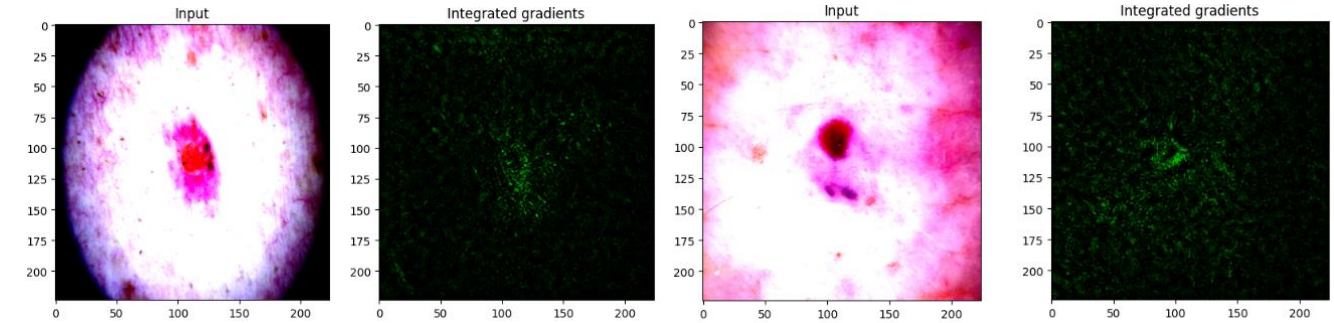


Figure 15: Some examples of lesions and associated masks from LIME, showing the important features. All images are of the top label output of the model.\

Integrated Gradients Examples



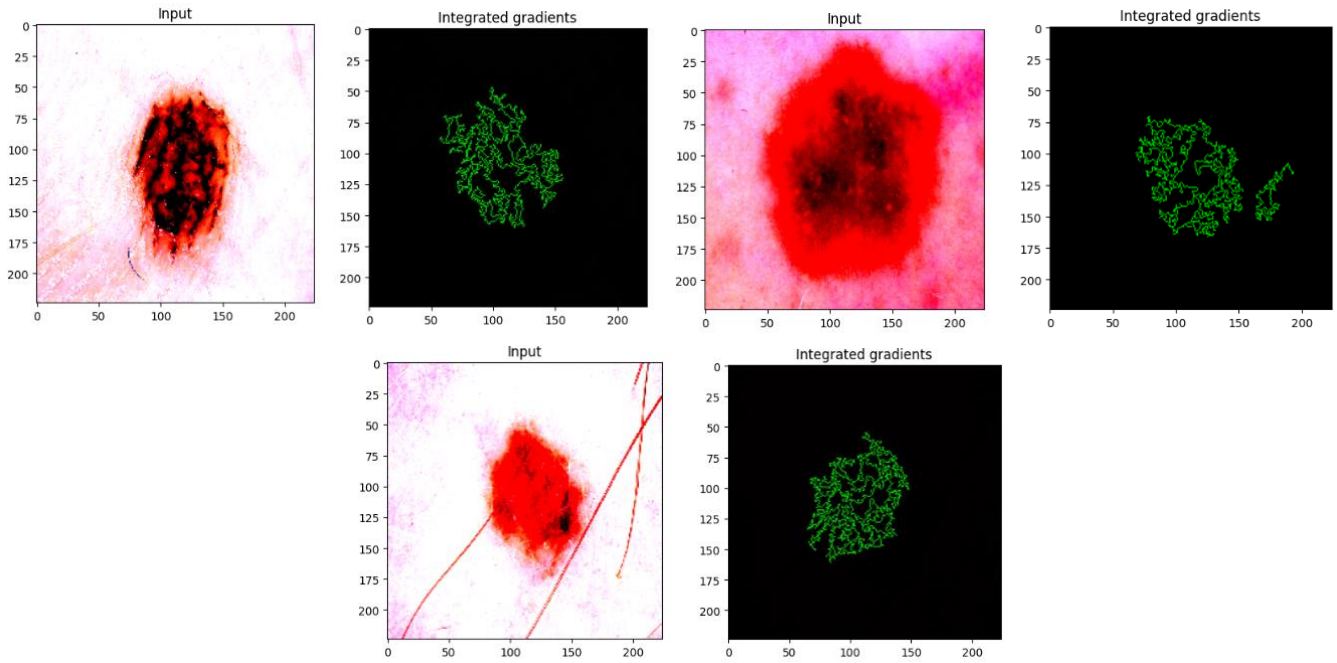


Figure 16: Some examples of Integrated Gradients showing relevant important features for the prediction.

6.6 Explanation Evaluations

Metric	Steps=20, Runs=2	Steps=20, Runs=20	Steps=50, Runs=2	Steps=50, Runs=20	Steps=80, Runs=2	Steps=150, Runs=2
Monotonicity Estimate	0.19	0.2114	0.2119	0.23	0.2263	0.2364
IROF (Input Reduction Output Fidelity)	62.17	64.641	64.31	64.62	62.28	64.29
Faithfulness Estimate	0.0312	0.0566	0.0281	0.0235	-	0.0208
Sparseness	0.4825	0.455	0.4785	0.4498	0.4786	0.459
Complexity	5.014	5.0602	5.0219	5.0702	5.0212	5.0549
Relative Output Stability	9.83E+09	6.34E+09	2.20E+10	1.56E+09	9.78E+08	2.05E+09
Relative Input Stability	8.96E+07	3.65E+07	1.68E+08	1.01E+07	5.84E+06	8.63E+06

Table 6: Evaluation of integrated gradients under different hyperparameters

All metrics were calculated and averaged over the same batch of 32 lesion images. Due to the computing/time constraints involved, a more involved analysis with further variation in the hyperparameters

and additional explainers could not be conducted. But there are still some preliminary insights to be gained from the data above. For example, both relative input and output stability go down as the number of runs are increased. This makes sense since the explainers can average out unimportant noisy features over multiple runs, leading to a better/lower stability score. On the other hand, a clear relationship between number of steps in integrated gradients and stability is not as apparent, although there is an improvement in stability as steps increase. This could possibly be due to the local 'stability' minima lying between some values. Of course, the lack of enough data is another possibility for the lack of an apparent relationship over this variable. Sparseness (lower is better) corroborates the insights from stability, suggesting that as more runs are allowed and more parameters made available to increase explainer complexity, it homes in on the most important features.

Complexity seems to increase as the number of steps, or the number of runs is increased. This could suggest the use of this metric to avoid overfitting due to larger hyperparameters that increase model complexity, but only in tandem with another metric that makes sure the explainer isn't underfit to the model (an explainer that doesn't do anything would achieve a complexity of zero for example). Monotonicity (higher is better) is another metric that improves as steps or runs are increased. Faithfulness Estimate and IROF (Input Reduction Output Fidelity) are both metrics that fluctuate over the tested space of hyperparameters, making it hard to gain insights from them. It is not clear whether this is an issue with sample space under which they were tested, an issue with the implementation of the calculation itself or whether these metrics might not be suitable for this specific task. It is worth noting that there were also other tested metrics, such as monotonicity, sufficiency and completeness that gave the same results for all combinations (All True/All False for instance) and were therefore not included in the table for differentiation purposes. From this analysis, two things become clear: there are clearly insights to be drawn as to which explainer can perform best by analysing them under different metrics but work still needs to be done to figure out which metrics to use and how to use them in tandem.

Chapter 7: Conclusion

This project focused on addressing the critical issue of evaluating explainability in AI-driven medical diagnostics, particularly within the context of a heavily class imbalanced skin cancer dataset. The primary hypothesis was that enhancing explainability through the evaluation of advanced post hoc methods would lead to improved trustworthiness and usability of AI models in medical diagnosis. The problem was modelled using state-of-the-art CNNs combined with transfer learning techniques, such as DenseNet and MobileNet. The approach involved a comprehensive evaluation of explanation methods like Integrated Gradients and Grad-CAM to assess their effectiveness in providing clear, faithful explanations that align with clinical requirements. The results of this study highlight both the promise and the challenges of current xAI methods. While some evaluation methods did indicate a potential correlation between the metrics and hyperparameter adjustments, these correlations were not consistent across all metrics, making it challenging to pinpoint which metrics should be prioritized in a standardized evaluation framework. It is possible that such a framework may need to be tailored to specific problems. Additionally, the current strategy of applying post hoc methods after model training might benefit from being integrated into the training process itself. By incorporating metrics like faithfulness and robustness during training, it may be possible to guide models toward more interpretable and reliable explanations, optimizing these aspects as part of the model's learning process rather than evaluating them after the fact.

Despite these promising results, several limitations remain that could impact the broader applicability and effectiveness of the proposed methods. One significant limitation is the reliance on pixel-based explanations, which may not capture the process in a way that is meaningful to clinicians. Future work could explore the use of segmentation-based assessments to provide more localized and interpretable explanations, like how radiologists assess regions of interest in medical images. Additionally, user studies involving healthcare professionals could be integrated into the evaluation process to ensure that the explanations generated by AI models align with clinical reasoning and practical decision-making. Another limitation is the presence of noise in the images, such as hair, which could be addressed by pre-processing steps like hair removal to improve the clarity of explanations. Furthermore, the current methods did not incorporate concept-based explanations that mirror diagnostic checklists used by doctors (e.g., ABCDE for melanoma assessment), which could enhance the relevance and interpretability of AI decisions. Future research could focus on developing and testing concept-based xAI methods that resonate more closely with clinical practices, as well as refining the evaluation metrics to include both qualitative and quantitative assessments from end-users. Additionally, extending the study to include diverse datasets and real-time, continuous learning models could further validate the robustness and generalizability of these methods.

References

A. A. Adegun and S. Viriri, "FCN-Based DenseNet Framework for Automated Detection and Classification of Skin Lesions in Dermoscopy Images," in *IEEE Access*, vol. 8, pp. 150377-150396, 2020, doi: 10.1109/ACCESS.2020.3016651.

Alfi, I.A.; Rahman, M.M.; Shorfuzzaman, M.; Nazir, A. A Non-Invasive Interpretable Diagnosis of Melanoma Skin Cancer Using Deep Learning and Ensemble Stacking of Machine Learning Models. *Diagnostics* **2022**, *12*, 726. https://doi.org/10.3390/diagnostics12030726

Ali, A.-R., Li, J. and O'Shea, S.J. (2020). Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images. *PLOS ONE*, 15(6), p.e0234352. doi:https://doi.org/10.1371/journal.pone.0234352.

Alvarez-Melis, David and Jaakkola, Tommi S.. "On the Robustness of Interpretability Methods." (2018)

Amann, J., Blasimme, A., Vayena, E. *et al.* Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* **20**, 310 (2020). https://doi.org/10.1186/s12911-020-01332-6

Australian Institute of Health and Welfare (2023). *Cancer Data in Australia, Overview of Cancer in Australia, 2023*. [online]. Available at: https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/overview-of-cancer-in-australia-2023

Binder, A., Bockmayr, M., Hägele, M. *et al.* Morphological and molecular breast cancer profiling through explainable machine learning. *Nat Mach Intell* **3**, 355–366 (2021). https://doi.org/10.1038/s42256-021-00303-4

Bhatt, U., Adrian Weller, and José M. F. Moura. 2021. Evaluating and aggregating feature-based model explanations. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20). Article 417, 3016–3022.

Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, Schilling B, Haferkamp S, Schadendorf D, Holland-Letz T, Utikal JS, von Kalle C; Collaborators. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer. 2019 May;113:47-54. doi: 10.1016/j.ejca.2019.04.001. Epub 2019 Apr 10. PMID: 30981091.

Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., & Esser, S. (2018). Skin cancer classification using convolutional neural networks: systematic review. *Journal of Medical Internet Research*, 21(10), e11936.

Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7

Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2023. Explainable Deep Learning Methods in Medical Image Classification: A Survey. ACM Comput. Surv. 56, 4, Article 85 (April 2024), 41 pages. https://doi.org/10.1145/3625287

C. Pandey, A. D. Choudhury and S. Khandelwal, "qxAI: Quantifiable xAI for Cardiac Diseases," *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Biarritz, France, 2024, pp. 233-238, doi: 10.1109/PerComWorkshops59983.2024.10502886.

Daneshjou R, Barata C, Betz-Stablein B, et al. Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology: CLEAR Derm Consensus Guidelines From the International Skin

Imaging Collaboration Artificial Intelligence Working Group. *JAMA Dermatol.* 2022;158(1):90–96. doi:10.1001/jamadermatol.2021.4915

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. doi: https://doi.org/10.48550/arXiv.1702.08608

Ehsan, U., Harrison, B., Chan, L. and Riedl, M.O. (2017). *Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations*. [online] arXiv.org. doi: https://doi.org/10.48550/arXiv.1702.07826.

Eloise Withnell, Xiaoyu Zhang, Kai Sun, Yike Guo, XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data, *Briefings in Bioinformatics*, Volume 22, Issue 6, November 2021, bbab315, https://doi.org/10.1093/bib/bbab315

Esteva, A., Kuprel, B., Novoa, R. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017). https://doi.org/10.1038/nature21056

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *COM/2021/206 final*.

Fong, R. and Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. doi: https://doi.org/10.1109/iccv.2017.371

Ghanvatkar, S., and Rajan, V. "Evaluating Explanations From AI Algorithms for Clinical Decision-Making: A Social Science-Based Approach," in *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 4269-4280, July 2024, doi: 10.1109/JBHI.2024.3393719

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gloster, H.M. and Neal, K. (2006). Skin cancer in skin of color. *Journal of the American Academy of Dermatology*, [online] 55(5), pp.741–760. doi: https://doi.org/10.1016/j.jaad.2005.08.063.

Qi Han, Xin Qian, Hongxiang Xu, Kepeng Wu, Lun Meng, Zicheng Qiu, Tengfei Weng, Baoping Zhou, Xianqiang Gao, DM-CNN: Dynamic Multi-scale Convolutional Neural Network with uncertainty quantification for medical image classification, Computers in Biology and Medicine, Volume 168, 2024, 107758, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2023.107758.

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, Uhlmann L; Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018 Aug 1;29(8):1836-1842. doi: 10.1093/annonc/mdy166. PMID: 29846502.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. doi: https://doi.org/10.48550/arXiv.1512.03385

Hedström, A., Weber, L., Bareeva, D., Krakowczyk, D., Motzkus, F., Samek, W., Lapuschkin, S. and Höhne, C. (2023). Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, [online] 24, pp.1–11. Available at: https://jmlr.org/papers/volume24/22-0142/22-0142.pdf [Accessed 13 Aug. 2024].

Hicks, S.A., Strümke, I., Thambawita, V. *et al.* On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* **12**, 5979 (2022). https://doi.org/10.1038/s41598-022-09954-8

Hossain MM, Hossain MM, Arefin MB, Akhtar F, Blake J. Combining State-of-the-Art Pre-Trained Deep Learning Models: A Noble Approach for Skin Cancer Detection Using Max Voting Ensemble. Diagnostics (Basel). 2023 Dec 30;14(1):89. doi: 10.3390/diagnostics14010089. PMID: 38201399; PMCID: PMC10795598.

Hu, B., B. Vasu and A. Hoogs, "X-MIR: EXplainable Medical Image Retrieval," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2022, pp. 1544-1554, doi: 10.1109/WACV51458.2022.00161.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. doi: https://doi.org/10.48550/arXiv.1608.06993

Huang, H.-W., Hsu, B.W.-Y., Lee, C.-H. and Tseng, V.S. (2021), Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers. J. Dermatol., 48: 310-316. https://doi.org/10.1111/1346-8138.15683

Holzinger, A., Biemann, C., Pattichis, C. and Kell, D. (2017). *What do we need to build explainable AI systems for the medical domain?* [online] Available at: https://arxiv.org/pdf/1712.09923.pdf.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. doi: https://doi.org/10.48550/arXiv.1704.04861

International Skin Imaging Collaboration. *SLICE-3D 2024 Challenge Dataset. International Skin Imaging Collaboration https://doi.org/10.34970/2024-slice-3d (2024).*

International Organization for Standardization (ISO). (2016). ISO 13485:2016 Medical devices – Quality management systems – Requirements for regulatory purposes.

International Organization for Standardization (ISO). (2019). ISO 14971:2019 Medical devices – Application of risk management to medical devices.

International Organization for Standardization (ISO). (2015). ISO 14001:2015 Environmental management systems – Requirements with guidance for use.

International Organization for Standardization (ISO). (2008). ISO 9241-171:2008 Ergonomics of human-system interaction – Part 171: Guidance on software accessibility.

Jin, W., Li, X. and Hamarneh, G. (2022). Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11), pp.11945–11953. doi: https://doi.org/10.1609/aaai.v36i11.21452

Jin, W., Xiaoxiao Li, Mostafa Fatehi, Ghassan Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, Volume 84, 2023, 102684, ISSN 1361-8415, https://doi.org/10.1016/j.media.2022.102684

K. M. Hosny, M. A. Kassem and M. M. Foaud, "Skin Cancer Classification using Deep Learning and Transfer Learning," *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*, Cairo, Egypt, 2018, pp. 90-93, doi: 10.1109/CIBEC.2018.8641762.

Kumar Lilhore, U., Simaiya, S., Sharma, Y.K. *et al.* A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameters optimization. *Sci Rep* **14**, 4299 (2024). https://doi.org/10.1038/s41598-024-54212-8

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *doi:* https://doi.org/10.48550/arXiv.1412.6980

L. N. Smith, "Cyclical Learning Rates for Training Neural Networks," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 2017, pp. 464-472, doi: 10.1109/WACV.2017.58.

Ladbury, C., Reza Zarinshenas, Hemal Semwal, Tam, A., Nagarajan Vaidehi, Rodin, A.S., Liu, A., Glaser, S., Ravi Salgia and Amini, A. (2022). Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Transatlantic Cancer Research*, 11(10), pp.3853–3868. doi: https://doi.org/10.21037/tcr-22-1626.

Li, Y., Zhou, J., Verma, S. and Chen, F. (2023). *A Survey of Explainable Graph Neural Networks: Taxonomy and Evaluation Metrics*. [online] Available at: https://arxiv.org/pdf/2207.12599.pdf [Accessed 9 Feb. 2024].

Luis A. de Souza, Robert Mendel, Sophia Strasser, Alanna Ebigbo, Andreas Probst, Helmut Messmann, João P. Papa, Christoph Palm, Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box, Computers in Biology and Medicine, Volume 135, 2021, 104578, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104578

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. doi: https://doi.org/10.48550/arXiv.1705.07874

Md. Kamrul Hasan, Md. Toufick E. Elahi, Md. Ashraful Alam, Md. Tasnim Jawad, Robert Martí, DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation, Informatics in Medicine Unlocked, Volume 28, 2022, 100819, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100819.

Md. Aminur Rab Ratul, M. Hamed Mozaffari, Won-Sook Lee, Enea Parimbelli, Skin Lesions Classification Using Deep Learning Based on Dilated Convolution, bioRxiv 860700; doi: https://doi.org/10.1101/860700

M. Sobhan and A. M. Mondal, "Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 3152-3159, doi: 10.1109/BIBM55620.2022.9995516.

Magesh, P.R., Myloth, R.D. and Tom, R.J. (2020). An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. *Computers in Biology and Medicine*, 126, p.104041. doi: https://doi.org/10.1016/j.compbiomed.2020.104041

Makridis *et al*., "Towards a Unified Multidimensional Explainability Metric: Evaluating Trustworthiness in AI Models," *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, Pafos, Cyprus, 2023, pp. 504-511, doi: 10.1109/DCOSS-IoT58021.2023.00084.

Miller, T., Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Volume 267, 2019, Pages 1-38, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2018.07.007.

Mohan, H. and Yoo, J. (2023). A comprehensive analysis of recent advancements in cancer detection using machine learning and deep learning models for improved diagnostics. *Journal of Cancer Research and Clinical Oncology*, 149(15), pp.14365–14408. doi:https://doi.org/10.1007/s00432-023-05216-w.

Mukherjee, S., Adhikari, A. and Roy, M. (2019). Malignant Melanoma Classification Using Cross-Platform Dataset with Deep Learning CNN Architecture. *Recent Trends in Signal and Image Processing*, pp.31–41. doi:https://doi.org/10.1007/978-981-13-6783-0_4.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 1 (January 2002), 321–357.

Pieter Van Molle, Miguel De Strooper, Verbelen, T., Vankeirsbilck, B., Simoens, P. and Dhoedt, B. (2018). Visualizing Convolutional Neural Networks to Improve Decision Support for Skin Lesion Classification. *Lecture Notes in Computer Science*, pp.115–123. doi:https://doi.org/10.1007/978-3-030-02628-8_13.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W. and Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*. [online] Available at: https://arxiv.org/abs/1802.07810.

Priti Bansal, Ritik Garg, Priyank Soni, Detection of melanoma in dermoscopic images by integrating features extracted using handcrafted and deep learning models, Computers & Industrial Engineering, Volume 168, 2022, 108060, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2022.108060.

Rajendran, B., Simeone, O. and Al-Hashimi, B. (2023). *Towards Efficient and Trustworthy AI Through Hardware-Algorithm-Communication Co-Design*. [online] Available at: https://arxiv.org/pdf/2309.15942.pdf [Accessed 8 Feb. 2024].

Rezazadeh, A., Jafarian, Y. and Kord, A. (2022). Explainable Ensemble Machine Learning for Breast Cancer Diagnosis Based on Ultrasound Image Texture Features. *Forecasting*, 4(1), pp.262–274. doi: https://doi.org/10.3390/forecast4010015.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. doi: https://doi.org/10.48550/arXiv.1602.04938

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

Salih O, Duffy KJ. Optimization Convolutional Neural Network for Automatic Skin Lesion Diagnosis Using a Genetic Algorithm. *Applied Sciences*. 2023; 13(5):3248. https://doi.org/10.3390/app13053248

Sangwan H., 2024, Github Repository, https://github.com/hardikSangwan/thesis_diagnostics_skin

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. doi: https://doi.org/10.48550/arXiv.1610.02391

Shahsavari, A., Khatibi, T. & Ranjbari, S. Skin lesion detection using an ensemble of deep models: SLDED. *Multimed Tools Appl* **82**, 10575–10594 (2023). https://doi.org/10.1007/s11042-022-13666-6

Siddiqui, K. and T. E. Doyle, "Trust Metrics for Medical Deep Learning Using Explainable-AI Ensemble for Time Series Classification," *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Halifax, NS, Canada, 2022, pp. 370-377, doi: 10.1109/CCECE49351.2022.9918458.

Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *J. Imaging* **2020**, *6*, 52. https://doi.org/10.3390/jimaging6060052

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks, https://arxiv.org/abs/1703.01365

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. doi: https://doi.org/10.48550/arXiv.1602.07261

T. Alkarakatly, S. Eidhah, M. Al-Sarawani, A. Al-Sobhi and M. Bilal, "Skin Lesions Identification Using Deep Convolutional Neural Network," *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, Al

Tabrizchi, H., Parvizpour, S. & Razmara, J. An Improved VGG Model for Skin Cancer Detection. *Neural Process Lett* **55**, 3715–3732 (2023). https://doi.org/10.1007/s11063-022-10927-1

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15). Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

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

United Nations Sustainable Development. *Home - 2019 - United Nations Sustainable Development*. [online] Available at: https://www.un.org/sustainabledevelopment.

U.S. Food and Drug Administration (FDA). (2021). Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan.

Wang, C. (2023). *Calibration in Deep Learning: A Survey of the State-of-the-Art*. [online] Available at: https://arxiv.org/pdf/2308.01222.pdf [Accessed 8 Feb. 2024].

Xu, Y., Lam, H.-K. and Jia, G. (2021). MANet: A two-stage deep learning method for classification of COVID-19 from Chest X-ray images. *Neurocomputing*, 443, pp.96–105. doi: https://doi.org/10.1016/j.neucom.2021.03.034

Y. Bengio, Neural Networks: Tricks of the Trade chapter Practical recommendations for gradient-based training of deep architectures, Springer Berlin Heidelberg, pp. 437-478, 2012.

Y. Jiang, S. Cao, S. Tao and H. Zhang, "Skin Lesion Segmentation Based on Multi-Scale Attention Convolutional Neural Network," in *IEEE Access*, vol. 8, pp. 122811-122825, 2020, doi: 10.1109/ACCESS.2020.3007512.

Yaqoob, A., Rabia Musheer Aziz and Navneet Kumar Verma (2023). Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review. *Human-Centric Intelligent Systems*. doi: https://doi.org/10.1007/s44230-023-00041-3.

Yosinski, J., Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, 3320–3328.

Zhang, Li, Jian Zhang, Wenlian Gao, Fengfeng Bai, Nan Li, Noradin Ghadimi, A deep learning outline aimed at prompt skin cancer detection utilizing gated recurrent unit networks and improved orca predation algorithm, Biomedical Signal Processing and Control, Volume 90, 2024, 105858, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2023.105858.

Zhiwei Qin, Zhao Liu, Ping Zhu, Yongbo Xue, A GAN-based image synthesis method for skin lesion classification, Computer Methods and Programs in Biomedicine, Volume 195, 2020, 105568, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2020.105568.

Zou, L. *et al*., "Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections," in *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 2, pp. 242-254, April 2023, doi: 10.1109/TAI.2022.3153754