AI and Sensor Development for Skin Cancer Detection

EMS700U — Industry / Research Project — 2025/26

Individual Submission Title:

Dataset Curation and Machine Learning Model Development

Author:	Jatinder Dhaliwal (220451019)
Team Members:	Emily Speed
	Kennasa Ahmed
	Nadia Mohammad
	Gurtej Panesar
	Holly Azarinejad
Group Number:	244
Project Supervisors:	Zion Tse, Hadi Sadati
Module:	EMS700U — Industry / Research Project
Academic Year:	2025/26
Submission Date:	29 April 2026
Word Count:	7470

School of Engineering and Materials Science

Queen Mary University of London

Abstract

This project works to develop a mobile application to assist general practitioners (GPs) in screening patients for skin cancer, with the aim of improving NHS triaging by allowing an additional measure alongside relevant patient information. The back-end portion involved curating two large-scale datasets – an unclean and clean dataset – each including a total of 3680 benign and malignant skin lesion images, to train, validate, and test a convolutional neural network (CNN) model for classification of lesions. Simultaneously, the effects of strict quality control on dataset curation, and the optimisation of epoch in the machine learning model were investigated. The results showed that the model achieved the highest performance when trained on the unclean dataset at 25 epochs, with an overall accuracy of 84.8% compared to 79.4% on the clean dataset, considering various metrics including sensitivity, specificity, precision, and loss. This was attributed to the unclean dataset representing variation in real-world data closer than the clean dataset, hence improving generalisation and providing more accurate classification. The model was also compared to three other CNN architectures on the same datasets, where the model outperformed all CNNs. Overall, this project demonstrated the importance of curating a representative dataset, as well as the effect of epoch adjustment, on model performance. However, additional measures such as optimising more model parameters, obtaining data from international and private databases, and testing other deep learning algorithms, may be required to further examine how to optimise machine learning systems in dermatology applications.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviation	Definition
ABCDE	Asymmetry, Border, Colour, Diameter, Evolving
AI	Artificial Intelligence
AP	Average Precision
AUC	Area Under the Curve
CE	Conformité Européenne
CNN	Convolutional Neural Network
FDA	Food and Drug Administration
FNR	False Negative Rate
FPR	False Positive Rate
FST	Fitzpatrick Skin Type
GP	General Practitioner
GradCAM	Gradient-weighted Class Activation Mapping
HAM10000	Human Against Machine with 10,000 training images
ISIC	International Skin Imaging Collaboration
KNN	K-Nearest Neighbours
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NPV	Negative Predictive Value
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
UI	User Interface
UK	United Kingdom
UKCA	UK Conformity Assessed
US	United States
UX	User Experience
XAI	Explainable Artificial Intelligence

Declaration of Work

I declare that this dissertation is my own work and has not been submitted in any form for another degree or institution. All sources of information have been acknowledged and referenced appropriately.

This project was undertaken as part of a group project. However, the work presented in this dissertation represents my individual contribution. The contributions of other group members have not been presented as my own.

Signature:

Date: 29/04/2026

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Zion Tse, Brandon Ferral and Chayabhan Limpabandhu, for their guidance, feedback, and support throughout the course of this project.

I would also like to thank my group members, Gurtej Panesar, Nadia Mohammad, Kennasa Ahmed, Emily Speed, and Holly Azarinejad, for their collaboration during the project. Their contributions to the project framework were significant and helped shape the direction of this work.

1 Background

1.1 Project Rationale & Aims

In the UK, NHS dermatology services receive 1.2 million annual referrals. 60% are considered urgent, but only 6% are classified as malignant (NICE 2025). Additionally, there are currently 659 professional dermatologists in England (Levell 2021). The combination of this significantly strains the NHS, leading to delays in treatment and overshadowing of malignant cases. Furthermore, COVID-19 caused many cancellations of non-urgent referrals, resulting in a backlog of 441,000 patients, triggering an ongoing effect of making fast assessment increasingly difficult (NICE 2025).

With the intent to reduce unnecessary referrals and facilitate early treatment of melanoma, this project works to develop a machine learning model incorporated into a smartphone application to support general practitioners (GPs) in classifying skin lesions. By providing an additional measure alongside patient history, a GP can make an informed decision about whether a dermatologist referral is required. It should be noted that this application is a tool to assist GPs' decision-making rather than diagnosis; the concept being to provide supplementary information.

The rest of this report is designed as follows: The remainder of Section 1 explores the societal and ethical considerations that must be made, the objectives of the project, work allocation, including a work package flowchart, and the individual contribution to the project. Section 2 shows the literature review, looking at current methods and existing applications of lesion classification, a summary of the regulatory affairs associated with this project, deep learning algorithms, and model parameters to control. Section 3 displays the methodology of the individual contribution to this project. Section 4 includes the results, and Section 5 provides the subsequent analysis of outcomes. Section 6 discusses the implications of these results in clinical practice, and Section 7 presents additional factors to consider in future work.

1.2 Societal & Ethical Considerations

Many societal and ethical benefits are contingent on the performance and fairness of this app; incorrect results could become counterproductive and convert to unnecessary referrals or delay critical treatment. Such errors are attributed to a bias in the training of current technology. Dr. David Wen reviewed 21 publicly available databases and found 10 of 2436 images to be of brown skin and only one to be black (Wen et al. 2021). Another study was performed on two renowned dermatology algorithms, ModelDerm and DeepDerm, which showed a reduced sensitivity of Fitzpatrick Skin Type (FST) I-II and FST V-VI detection (Daneshjou et al. 2022). Furthermore, in the UK, less than 0.5% of skin cancer diagnoses are from black and Asian ethnic groups, increasing the challenge of collecting diverse data for a fair development of models (Ascott et al. 2023). Therefore, the data that is collected must aim to cover a range of skin tones to mitigate this issue. Figure 1 below shows a chart of FSTs I-VI.

Figure 1: *Fitzpatrick Skin Type (FST) scale. The skin tone gets progressively darker with a higher FST number. Image obtained from (PureLite 2021).*

1.3 Objectives & Work Package Flowchart

The objectives of this project are split into six individual workstreams, ensuring equal workload between team members. Each role assigned is pivotal to the project's structure and trajectory, encouraging regular communication and consistent task execution.

Many databases include lesion images with significant artefacts, decreasing the performance of computational systems, so the first role is to curate a clean image dataset that can be used to train a machine learning model. This should include a fairness monitoring stage, ensuring that the dataset covers all FSTs.

Although great care must be taken during the curation process, artefacts cannot be completely avoided. To offset this, a second role involves image processing, including pre-cleaning and segmentation techniques to remove artefacts, and enhancing image contrast, ensuring no external factors impact model analysis.

The next role is to develop a classification model to categorise skin lesions into benign and malignant classes, based on feature extraction. Initially, the model will produce binary outcomes but will be updated to accurately classify subtypes of benign lesions using multi-class classification. Fairness monitoring should also be implemented here, ensuring the model does not bias toward or against a specific FST.

Alongside the model, this application includes an explainable AI (XAI) output, providing reasoning for an outcome for each lesion image. Each result will be presented with a percentage likelihood of malignancy, as well as a model confidence score to show the difficulty in analysing complex images.

The final role in the back-end portion involves validation and identifying and minimising signs of bias. This would be followed by performance tracking in real-time, crucial for highlighting model drift, where the model accuracy gradually decreases due to changes in input and output data and relationships (IBM, 2024).

User interface (UI) and user experience (UX) development covers the front-end portion. The user requirements can provide insights into what design and information the application should include. Since the target audience is GPs, the app should be less tailored towards informative aspects and instead should focus on simplicity and ease of navigation.

Table 1 provides a summary of the distribution of work within the team, followed by Figure 2, which illustrates a flowchart of the individual workstreams contributing to the overall project.

Table 1: *Distribution of work for each team member in the project. There are six members in total, allowing for six separate workstreams and ensuring everyone in the project can have an equal impact on the project. The completion of each workstream is key to produce a finished application.*

Member	Workstream	Responsibilities
Gurtej	Ethical/Fairness Monitoring	Ensure even coverage of FSTs in the skin lesion dataset. Check whether model performance is consistent across all FSTs.
Emily	Image Processing	Involves pre-cleaning of images, segmentation techniques and enhancing image contrast to ensure no effect of external factors.
Jatinder	**Classification Model Development**	**Producing a classification model for detection of benign and malignant lesions. Developing and optimising the model through iterations and adjusting parameters.**
Kennasa	Explainable AI Output	Provides justifications for model outcomes, showing a model confidence score to determine complexity of analysis.
Holly	Performance Tracking/Validation	Validate and track model performance post-training on a range of skin lesion images, simulating real-world scenarios and highlighting potential model drift.
Nadia	UX Consideration and UI Design	Study the user requirements (GPs) and tailor the app interface to facilitate simplicity and a suitable user experience.

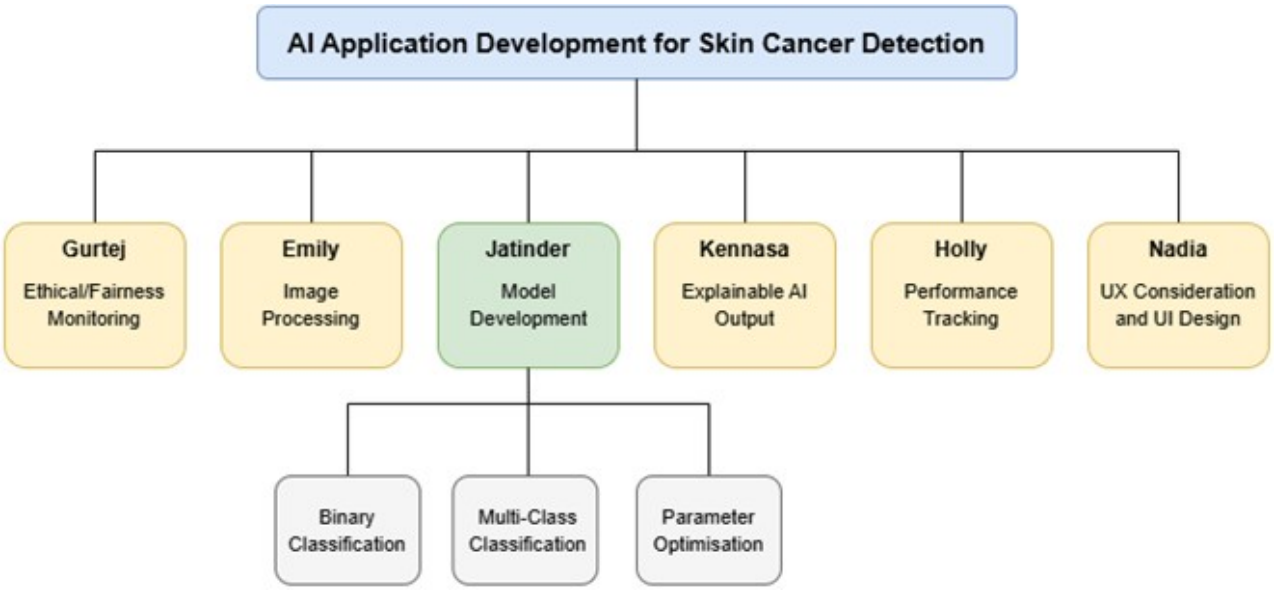

Figure 2: *Work package flowchart illustrating the six separate workstreams in this project. For this individual contribution, the three primary responsibilities have been listed underneath the workstream.*

1.4 Individual Contribution

1.4.1 Dataset Curation

Although the primary workstream is model development, this is heavily linked to the dataset curation process. The first objective is to collect a small dataset of images, practicing how to curate a clean dataset, locating sources with dermoscopic images, and creating a sample set to test the initial classification model. Using a low-scale dataset develops a baseline understanding of what is considered a high-quality image and allow errors to be outlined and resolved efficiently, instead of interpreting a large dataset immediately. Another purpose of this task is to self-analyse skin lesions to understand the criteria for benign and malignant classification.

Once the initial dataset has been run through a binary model, the next stage involves collecting a larger dataset to allow the model to train under more data. Simultaneously, the benign class must be split into multiple subtypes to test the model's ability to separate lesions in a multi-classification system. The justification behind selecting each subtype must be clear and relevant to common issues in dermatology. This process requires more scrutiny than the smaller dataset on minimising artefacts and balancing FSTs across images. When initially curating the dataset, some artefacts cannot be avoided, so this is referred to as the "unclean" dataset. The "clean" dataset has more stringency on removing images with artefacts, leaving clear lesion images. Figure 3 displays a clean image and an unclean image for direct comparison. After both datasets are created, they will be subsequently run through the machine learning model.

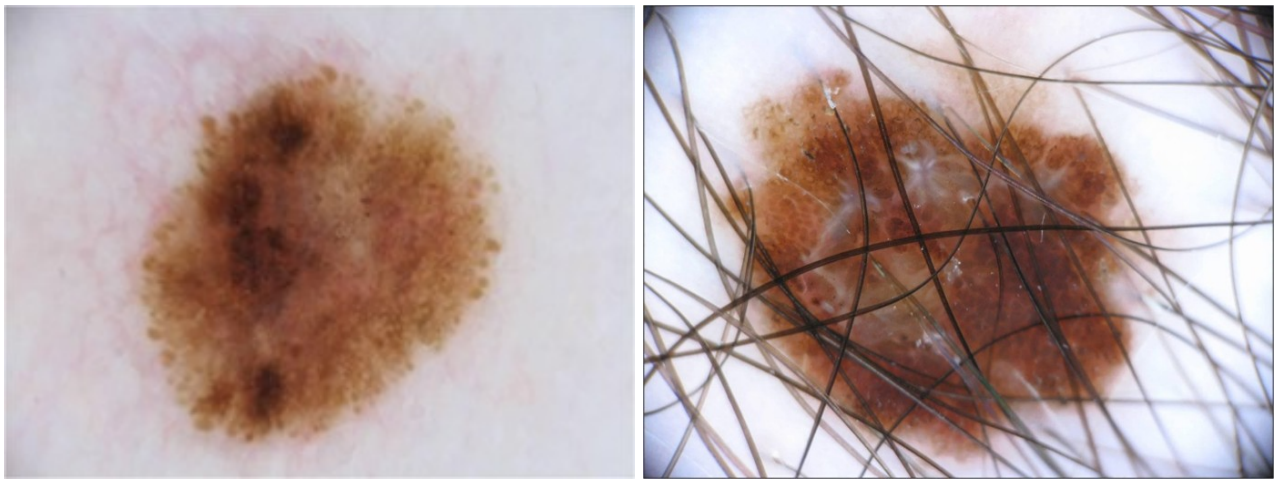

Figure 3: *(a) Clean image of skin lesion and (b) unclean image of skin lesion obtained from the ISIC Archive. The clean image lesion is unobstructed so no external artefacts are likely to be picked up by the model, whereas the unclean image has hairs covering the lesion which could influence the model outcome.*

1.4.2 Machine Learning Model Development

For the classification model, the first stage involves producing a binary outcome, showing either benign or malignant results for the smaller dataset. It is expected that the initial performance of the model would be inadequate due to various factors such as a lack of training data, the quality of training data, and problems with the parameters, producing a poor outcome. However, the model will need to be continuously refined and re-tested on both the small and large dataset to ensure it achieves this aim.

Once a larger dataset has been curated, the images can be run through an updated binary model with adjusted parameters based on previous performance, hence projecting a higher accuracy due to significantly broader datasets and a refined model setup. A similar size in both the unclean and clean datasets is critical to assess the impact of strict quality control during dataset curation on model performance. Once both datasets have been run through the model, the model structure can then be updated further to facilitate multiple classes.

2 Literature Review

2.1 Current Methods

The current method of skin cancer screening involves an initial assessment of skin lesions using the ABCDE criteria to classify lesions as benign or malignant, shown in Table 2. However, the subjectiveness of this task highlights a need for additional tools to aid in diagnosis. Many lesions referred to dermatologists are benign, highlighting the inefficiency of this approach (Tsao et al. 2015).

Table 2: *ABCDE criteria for lesion analysis (MASCED 2025). Skin lesions are categorised into either benign or malignant based on the features listed below, and those which warrant concern typically result in a secondary referral, although this is subjective to the dermatologist/GP.*

Criteria	**Benign Attributes**	**Malignant Attributes**
Asymmetry	Symmetrical in shape, typically round.	Asymmetry of shape.
Border	Well-defined, smooth borders.	Poorly defined or fading borders, presence of notching.
Colour	Single, uniform colour.	Multiple colours with uneven distribution.
Diameter	Smaller than 6 mm.	Larger than 6 mm.
Evolution	Appearance remains consistent over time.	Noticeable changes in size, colour, or border over weeks; may bleed or scab.

Existing applications show promise with a relatively high level of accuracy in identifying lesions. However, many tests performed using these tools are derived from controlled datasets and retrospective studies, so the reported accuracy may not be consistent when a random dataset is applied, raising concerns about the validity of these systems (Edge Health 2024). Table 3 displays the key features and limitations of two competing AI tools for dermatology: SkinVision and Skin Analytics (DERM). SkinVision and Skin Analytics function as "black box AI" models, meaning that outputs are generated without clear reasoning or uncertainty estimations (Skene et al. 2024), resulting in these applications being unsuitable for GP support.

Table 3: *Key features and limitations for two existing AI applications for dermatology use – SkinVision and Skin Analytics (DERM). Both are expensive and do not provide an explanation for specific outcomes of their respective models (black box output), hence being unsuitable for this project.*

Software	**Key Features**	**Limitations**
SkinVision	Provides an AI risk assessment. Conformité Européenne (CE) registered as Class I Medical Device.	Annual subscription cost of £49.99 (SkinVision 2025). Reported specificity of 80.1% (2022). Black-box AI output with limited interpretability.
Skin Analytics (DERM)	Classifies dermoscopic images into 11 lesion types. Used by GPs and dermatologists within the NHS. UKCA certified as a Class IIa device and CE marked as Class III (Skin Analytics 2023).	Cost of approximately £30 per referral plus additional fees. Reported specificity of 78.1% (NICE 2022). Limited dataset diversity across Fitzpatrick Skin Types (FST). Black-box AI output.

Skinive's application shows some promise, with their convolutional neural network (CNN) algorithm trained on 161,142 images and validated on 25,688 dermatologist-confirmed cases, which is exceptionally high and linked to higher precision. This has apparently led to 94.7% sensitivity in 2022; but may be inflated, as a vast proportion

of images were either benign or other diseases classes (Skinive 2022).

The main challenges of existing applications lie in diagnostic specificity, model transparency, and user trust (Skene et al. 2024). DERM is limited to post-GP referral use, facilitating triaging for dermatologists, unlike the aim of this project, which is to assist GP decision-making. In addition, it is difficult to compare model performance across studies when investigating metrics such as specificity, sensitivity, and negative predictive value (NPV).

2.2 Regulatory Affairs

The UKCA conformity mark must be given to any recognised medical device or software based in the UK (MHRA 2019). If the app is utilised internationally, then the corresponding regulations, including CE marking for the EU and FDA approval for the US, must be met. To adhere to these principles, it must be clear that the intended use is a tool for GPs to use alongside other relevant patient information. This prevents automation bias where medical professionals excessively rely on AI for a diagnostic result, and the use of XAI mitigates this (Goddard et al. 2012).

2.3 Deep Learning Algorithms

For lesion classification, the options for deep learning algorithms include CNNs, support vector machines (SVMs) and k-nearest neighbours (KNNs). CNNs are often used in classifying visual data due to their ability to identify spatial patterns such as edges and shapes. The hierarchical architecture allows for feature detection at many levels, allowing abstract features and details to be picked up. The number of parameters requiring training is significantly reduced with CNNs, hence improving generalisation of relationships between variables instead of memorisation (Indolia et al. 2018). As a result, CNNs rarely suffer from overfitting, where the model begins memorising noisy features (artefacts in this case), and fails to determine a general relationship. Underfitting occurs when the algorithm lacks sufficient training to identify the relationship between variables, although this is less attributed to issues with the neural network (Bashir et al. 2020).

A study performed in 2017 investigated the performance of a CNN model trained on 129,450 clinical images, significantly higher than previous datasets, on two binary classification cases – keratinocyte carcinomas versus seborrheic keratosis, and melanomas versus nevi. Performance was measured based on sensitivity and specificity, and the results from using a smaller dataset showed that the model outperformed 21 dermatologists at keratinocyte carcinoma and melanoma recognition. The values for sensitivity and specificity were consistent when subsequently tested on a large dataset, highlighting the robustness of CNNs in lesion classification (Esteva et al. 2017). A similar research study was conducted comparing a CNN against 58 dermatologists worldwide. The dermatologists were asked to diagnose lesions using purely the images (study level-I) and then with additional information (study level-II). As expected, the mean sensitivity increased from 86.6% at study level-I to 88.9% at level-II. Surprisingly, the CNN specificity values at those sensitivities outperformed most dermatologists, although 13 of 58 dermatologists outperformed the CNN. This indicates that CNNs are suitable tools to aid dermatologists in melanoma detection but also highlights the importance of patient context in skin cancer screening, and that dermatologists should consider management decisions alongside diagnostic classification (Haenssle et al. 2018).

SVMs have become increasingly popular for potential use on multi-classification systems. Additionally, they have been integrated with advanced methods including evolve algorithms to enhance classification ability and optimise parameters. The decision functions of SVMs are determined directly from training data in a way that separation of decision borders is maximised in a highly dimensional space, minimising classification errors and resulting in more generalisation by setting clearer boundaries between classes. SVMs have been used in many binary classifications and been successful in image detection and bioinformatics, which are relevant to this project. However, the limitations relate to algorithmic complexity, selection of parameters, and the performance

of SVMs on unbalanced datasets. As the dataset is increased, the computational process slows down significantly (Cervantes et al. 2020); in an NHS GP setting, where many images would be analysed, this system would likely fail. Also, the data in dermatology tends to bias towards lighter skin tones, and SVMs are commonly known for biasing for the minority class, therefore, decreasing accuracy on darker skin tones and further contributing to the unfairness of current AI models.

KNN classifiers are typically used in pattern recognition applications and are basic, non-parametric models that can be implemented into many systems. They work by determining characteristics of an object and then assigning it to a class with the minimum average distance to it. In skin lesion classification, this could include the size, shape or colour of a lesion, and classifying as benign or malignant, or specific subtypes of both, based on what the lesion looks the most likely to be. Although conceptually simple, KNNs have highly complex calculation methods, and consequently, only perform efficiently on small datasets due to smaller distances between neighbours. Additionally, KNN algorithms are fully dependent on the training dataset, emphasising a need for specifically selected images which can be time consuming (Kataria & Singh 2013). A study in 2016 analysed the performance of KNNs on discrete, continuous and combined medical datasets. It was found that the KNN classifier performed well over categorical and numerical datasets but struggled with the mixed dataset. Although this project involves categorical data, KNN algorithms may be unsuitable due to limited function on larger datasets, which is imperative for maximising training time and reliability (Hu et al. 2016).

2.4 Model Parameters

Finally, the effects of various parameters on model performance were reviewed. Image size can highly influence the image clarity and computational efficiency. A study performed in 2023 found that an image size of 128x128 pixels took half of the training time compared to 256x256 pixels. Although the former was outperformed by the latter, there was only a marginal difference observed, suggesting that for most purposes, 128x128 pixels is ideal for a balance between image quality and run time (Rukundo 2023).

Batch size is a parameter that controls the number of training images the model processes at once before updating weights. 32 is the default value in deep learning; anything higher speeds up computation but decreases updates which could prevent the network from converging sufficiently, whereas anything lower would significantly increase convergence time (Kandel & Castelli 2020).

Epoch refers to one pass over the entire dataset, and optimisation of this value depends on two factors – overfitting and underfitting. Continually adjusting the epoch value can control the model performance to an extent by determining the level of overfitting/underfitting and can be optimised to increase the model's generalisation (Hussein & Shareef 2024).

Convolutional blocks are introduced to models to enhance feature extraction, particularly in blurry images (Liu et al. 2024). In combination with filters, these can be used to identify specific visual features. Regarding skin lesions, these include colour, texture, borders, and complex features associated with malignancy. Each filter size corresponds to a particular visual feature, and as the filter size is increased, the characteristic being detected increases in complexity. To balance computational expense and detailed analysis of images, the number of convolutional blocks and filter sizes are minimised to each application's needs.

Kernel size determines how large an area of the image is examined at once. As the number of kernels is increased, the computation time increases due to the high number of parameters being updated (Öztürk et al. 2018). Therefore, the common kernel size in CNNs is 3x3 pixels, increasing accuracy whilst maximising operational efficiency.

Normalising input data of neural networks to a consistent scale is beneficial towards training a model. In deep learning, batch normalisation allows a constant standard deviation across intermediate layers, allowing for more stability in output data (Bjorck et al. 2018). As a result, this can be applied to maintain a relatively high level

of analytical detail without compromising model stability across skin lesions, hence increasing the robustness of the model.

The dropout value of a deep network contributes to preventing overfitting by randomly dropping units from the neural network during training (Srivastava et al. 2014). This stops the model from relying on and memorising a specific dataset and instead changes the focus to identifying general relationships and attributing relevant features to each class. A dropout value of 0.5 is suitable for most models, allowing sufficient units to identify trends while avoiding significant attachment to noise and artefacts.

3 Methodology

3.1 Dataset Curation

The first step in data collection involved curating a small-scale dataset composed of approximately 200 images, aiming for an equal number of benign and malignant images. These were found using databases such as DermNet and Vita Medical to obtain dermoscopic images to train the model on a combination of suspicious and clearly classified lesions. It was important to select images with minimal artefacts as these could negatively impact the model performance by obstructing the lesions or drawing the model's attention elsewhere. Artefacts include hairs and other non-important details on the patient's skin, poor lighting and blurriness in the image, or simply an unclear contrast between the patient's skin and the lesion itself. Although these could be mitigated using segmentation techniques and image processing, it was good practice to minimise these obstructions in the collection process and allowed the model to be tested early in the project, without the need to manually remove artefacts.

This task also served as an opportunity to identify concerning features of a skin lesion and perform self-analysis on some images using the ABCDE criteria. Although it was simple to notice details such as asymmetry, border and colour gradients, measuring diameter was quite difficult as few images had reference scales present to represent the lesion size more accurately. Additionally, the reference scales could potentially be considered an artefact as they are visible in an image taken by a dermatoscope; however, at this point in the project, these images were included for the purpose of the exercise. The final criterion, evolution, was not considered during self-analysis, as from a singular image, it was impossible to tell how a lesion would progress over time without extensive knowledge of specific skin lesion features. Overall, this provided a solid foundation of knowledge of benign and malignant attributes and how to extract features from an image, both of which translated effectively towards creating the machine learning model and evaluating XAI output later in the project.

The next stage of the dataset curation process involved creating a larger dataset, consisting of approximately 6000 images in total, which were split up into equal parts benign and malignant images. Additionally, the benign images were further divided into four subtypes with 750 images each. Three subtypes were proposed by supervisors – Intradermal Nevus, Seborrheic Keratosis and Cherry Angioma. These were chosen as they are among the most misdiagnosed benign lesions in clinical practice, particularly in primary care (Patient.info 2022). Solar Lentigo was the final subtype selected due to its prevalence in dermatology settings, hence a wide availability of images. It was found that there were not enough Cherry Angioma images available in public databases; to maintain equality across subtypes, it was replaced with the Dermatofibroma subtype due to its abundance, as well as being commonly misdiagnosed. Once the images were obtained, they were organised into folders of benign subtypes and malignant lesions. In addition, the images were divided into training, validation, and test data, with a 70%, 20% and 10% split, respectively.

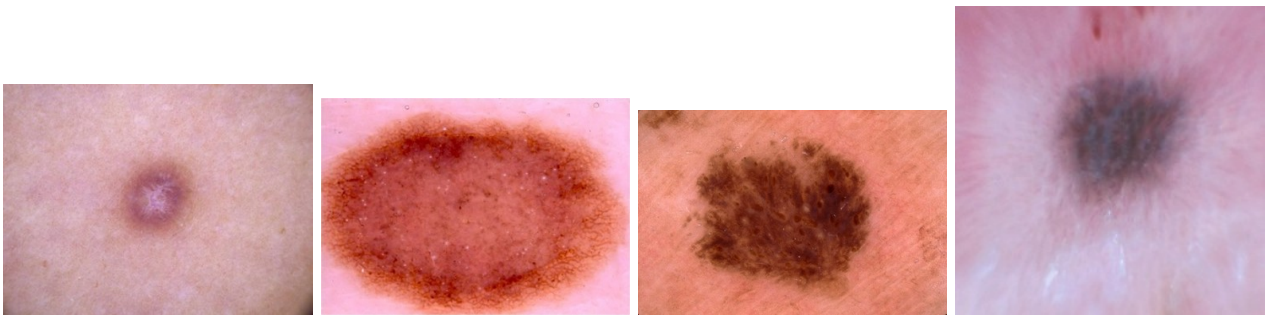

Figure 4: *Example images of (a) Dermatofibroma, (b) Intradermal Nevus, (c) Seborrheic Keratosis, and (d) Solar Lentigo, which were the four benign subtypes used in the dataset due to common misdiagnosis and wide availability of images.*

When curating the dataset, substantial databases such as The International Skin Imaging Collaboration (ISIC) Archive were used, which is a public database with 552,598 images in total (ISIC Archive n.d.), alongside HAM10000 (Kaggle), and DermNet which was used previously. This made it significantly easier to collect most images, although care had to be taken in avoiding duplicate images across databases. Many images showed artefacts which were filtered through, but some could not be avoided. Furthermore, the diversity of data across FSTs was extremely poor, with FST I-IV dominating the dataset. Though this was a primary aim of this project, due to time constraints and specific subtype requirements, it proved very difficult to collect an even distribution of images with different skin tones. Finally, the dataset was checked over, and any significantly obstructed images were removed from the dataset, reducing the total number to 3680, but, in return, creating a usable dataset which was referred to as the "unclean" dataset. Some images were replaced with cleaner images to create a "clean" dataset, and both were run through the binary classification model. The significant reduction in images meant that the benign lesions could not be split into subtypes, otherwise the model could not be sufficiently trained for each, hence lowering performance. The aim of the dataset curation process was to understand the implications of performing strict quality control on model performance in clinical practice. Table 4 shows the ratios of benign to malignant images in the datasets for each FST.

Table 4: *Ratio of benign to malignant images collected across each Fitzpatrick Skin Type (FST). FST V has an extremely high ratio, indicating major discrepancies in finding an equal number of images across skin types, especially malignant lesions.*

FST	Benign	Malignant	Ratio (B:M)
I	2668	244	11:1
II	3442	1445	2:1
III	1307	357	4:1
IV	878	32	27:1
V	815	5	163:1
VI	751	10	75:1

3.2 Model Development & Training

During the model development phase, Python was the software of choice, alongside PyCharm as the IDE due to efficiency in organising data and ability to share code. The deep learning algorithm chosen was CNN due to excellent classification and feature extraction capacity, and ease of parameter adjustments. For the initial dataset of 200 images, a template script of a binary machine learning model was provided, from which the image paths were edited before running the model. The purpose of this was to understand how a classification model is trained and validated and identify errors leading to false positives or false negatives. As expected, the

model performance was quite poor, predicting malignant for all cases; this is likely attributed to the lack of data, with only 156 and 38 images for training and validation, respectively, meaning that the model could not be sufficiently trained to distinguish between benign and malignant features. Therefore, a CNN algorithm was created with the assistance of LLMs and each parameter was justified. Finally, the CNN was run on the larger dataset. Table 5 shows the parameters and their justifications to optimise model performance.

Table 5: *Justification of parameters selected for the CNN model. Each of these have a significant impact specifically on model run time, computational expense, and accuracy in identifying features in skin lesion images.*

Parameter	**Justification**
Image Size (128×128 pixels)	Provides a balance between computational efficiency and image clarity, while preserving fine image details necessary for lesion analysis.
Batch Size (32)	Represents the number of training samples processed before weight updates. A batch size of 32 is widely used in classification tasks, offering stable performance without excessive computational demand.
Optimal Epoch Value (25)	One epoch corresponds to a full pass through the dataset. At 25 epochs, the model demonstrates sufficient learning without significant overfitting or underfitting.
Number of Convolutional Blocks (4)	Determines the depth of feature extraction. Four blocks enable hierarchical learning: early layers capture edges and colour gradients, intermediate layers detect textures and shapes, and deeper layers identify complex lesion patterns.
Filter Sizes (32, 64, 128, 256)	Each filter detects specific visual features. Increasing the number of filters across layers allows the model to capture progressively more complex representations.
Kernel Size (3×3 pixels)	Defines the receptive field of each filter. A 3×3 kernel is standard in CNNs, maintaining high accuracy while ensuring computational efficiency.
Batch Normalisation	Normalises layer outputs to maintain consistent value distributions, improving training stability and convergence speed.
Dropout (0.5)	Reduces overfitting by randomly deactivating 50% of neurons during training, encouraging the model to learn generalised patterns rather than memorising the dataset.

The sequential CNN used the Keras ImageDataGenerator to preprocess and feed images during training. The images from the training dataset were rescaled and normalised, and data augmentation was applied to rotate, zoom, and flip the image to facilitate generalisation. Data augmentation was not applied to the validation split, giving a fair evaluation of the model on unchanged images. Images were loaded in batches of 32 and resized to 128x128 pixels, with binary labels for benign and malignant lesions. The network had four convolutional blocks, each performing feature extraction, stabilising learning across layers using batch normalisation, and reducing spatial size, increasing efficiency and minimising computational expense. These blocks had filter sizes of 32, 64, 128 and 256, and as the filters increased, the complexity of features increased proportionally. This model extracted edges and textures from early layers, shapes and patterns from the intermediate layers, and object parts from deeper layers. A kernel size of 3x3 pixels was selected to capture finer details. Each layer was then flattened and dense with 256 neurons (units), with a dropout value of 0.5 applied, reducing the likelihood of overfitting. The output layer included one neuron with sigmoid activation, which outputs a probability and rounds up or down to the first class (benign) or the second class (malignant). The model was then assigned an epoch which was varied to find the optimal value, and the performance of the model was evaluated on the

validation data after each epoch, with weights being updated consistently during training. Finally, the model was saved in a .keras format and the training results were saved for further analysis of model performance. Figure 5 displays a flowchart illustrating the CNN model from input data to output result. Finally, due to limited time and resources for model training, the multi-classification system could not provide valid results; however, it was still attempted to understand the differences between binary and multi-class classification.

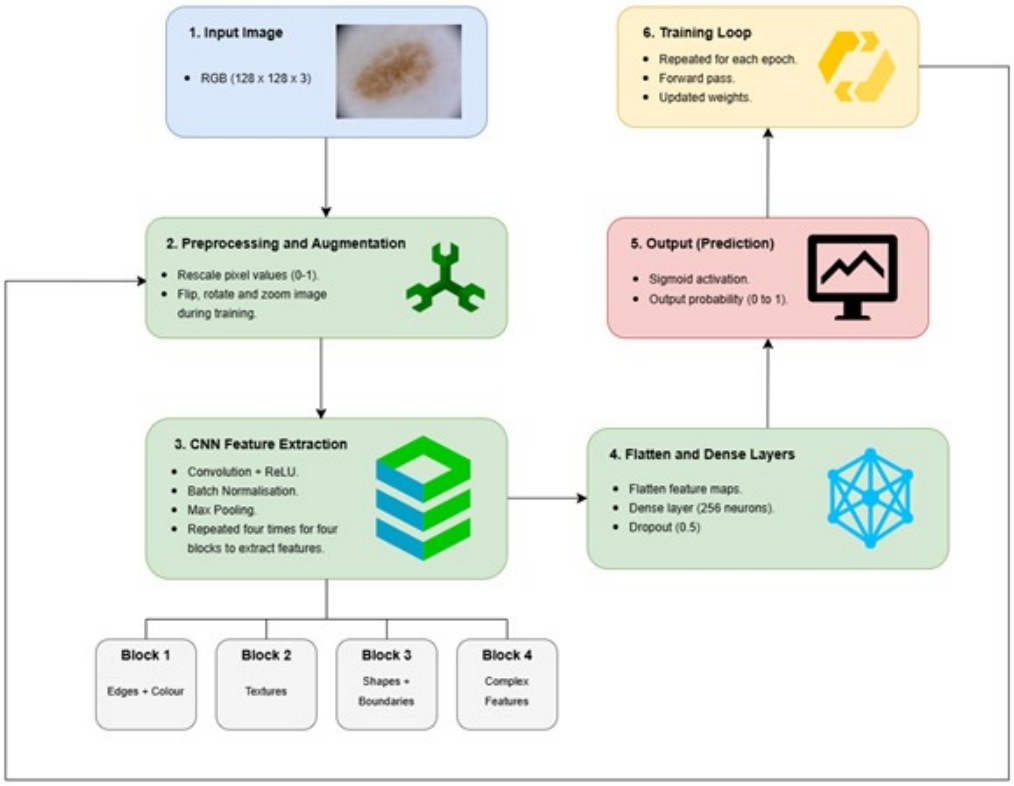

Figure 5: *Flowchart illustrating how the CNN model processes and outputs a result. This is repeated for each image in a training loop to ensure that the model has sufficient time to learn relationships associated with benign and malignant lesions.*

4 Results

4.1 Optimal Epoch

The epoch value was varied and run multiple times to define the optimal value for this classification model. The code was run from 0 to 25 epochs to find the point at which the model's accuracy peaked. Anything above 25 epochs took excessively long to compute, so 25 epochs was the maximum used. The optimal epoch was decided based on the accuracy and loss during validation, where the highest accuracy and lowest loss represented the best epoch for the model for both the unclean and clean datasets. Accuracy is the proportion of total predictions that the model correctly measured, whereas loss tracks the degree of error in the model's output (Bergmann and Stryker, 2024). Figure 6 displays accuracy and loss against epoch for both datasets, showing the effect of epoch on validation accuracy and loss. An epoch of 10 for the unclean dataset resulted in the highest validation accuracies of 82.47% and the lowest loss values at 0.3995, which were likely anomalous results. 25 epochs showed the highest validation accuracy of 81.76% and lowest loss at 0.3957 for the clean dataset.

Figure 6: *Graphs showing (a) validation accuracy against epoch and (b) validation loss against epoch to select optimal epoch value, for unclean and clean datasets. Both datasets show a very similar value for accuracy and loss at 25 epochs, with the unclean dataset slightly outperforming in both metrics.*

4.2 Sensitivity, Specificity & ROC Curve

Before looking at the following metrics, it should be noted that they are partly dependent on epoch, so where possible, the results at all epochs are shown, but for simplicity, the results of some metrics are shown at only 25 epochs in the main body, with the remainder available in Appendices. When comparing the results for the unclean and clean datasets, the principle metrics measured were sensitivity, specificity, accuracy, and precision after every five epochs up to 25. Sensitivity is the model's ability to correctly identify positive cases, whereas specificity is the ability to correctly identify negative cases. Positive cases refer to malignant lesions, and negative cases refer to benign lesions. Figure 7 illustrates bar charts of sensitivity and specificity for each dataset from 10 to 25 epochs, in intervals of five. The highest sensitivity in the unclean dataset was 95.98% at 10 epochs and in the clean dataset, 69.73% at 10 epochs, which were considered outliers. For specificity, the highest was noted at 78.92% at 15 epochs in the unclean dataset and 90.58% at 20 epochs in the clean dataset.

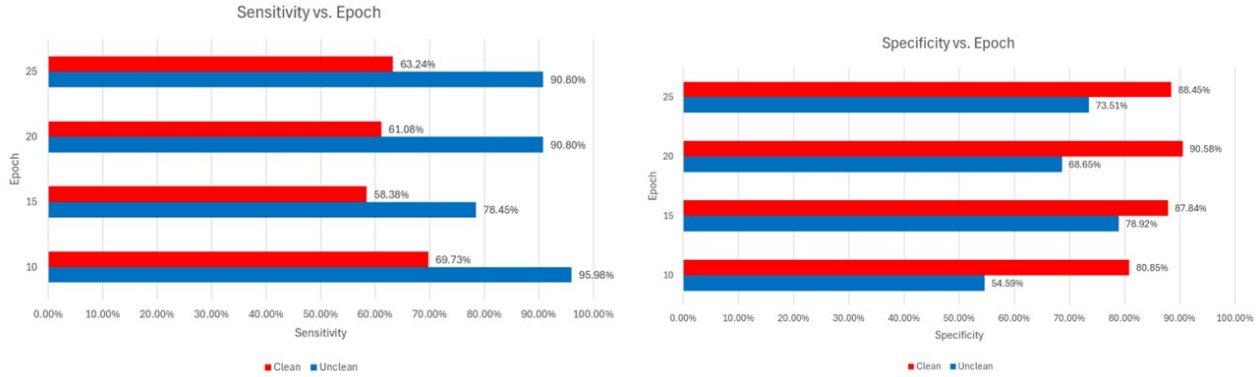

Figure 7: *Graphs showing (a) sensitivity against epoch and (b) specificity against epoch for unclean and clean datasets. At 10 epochs, a significantly higher sensitivity is seen, which is likely an anomalous result. At all epochs, the unclean dataset outperformed the clean dataset in terms of sensitivity, and vice versa for specificity.*

A Receiver Operating Characteristic (ROC) curve is based on the true positive rate (sensitivity) against the false positive rate (1 – specificity). The aim is to evaluate the performance of a binary classification model by accounting for the classification threshold of the model. A stricter criterion results in the curve shifting downwards and to the left, whereas a looser criterion shifts the curve upwards and to the right. An ROC curve can be summarised by looking at the area under the curve (AUC), which has a maximum value of one. A curve close to the top left corner has a higher AUC, indicating superior performance at all thresholds (Nahm 2022). Figure 8 shows the ROC curves at 25 epochs for both datasets. Both datasets have a similar curve, indicating little difference; however, the unclean dataset slightly outperformed the clean dataset, with an AUC of 0.9091

compared to 0.8934. The shapes of the curves are indicative of a strong ROC curve, being close to the top-left corner of the graph.

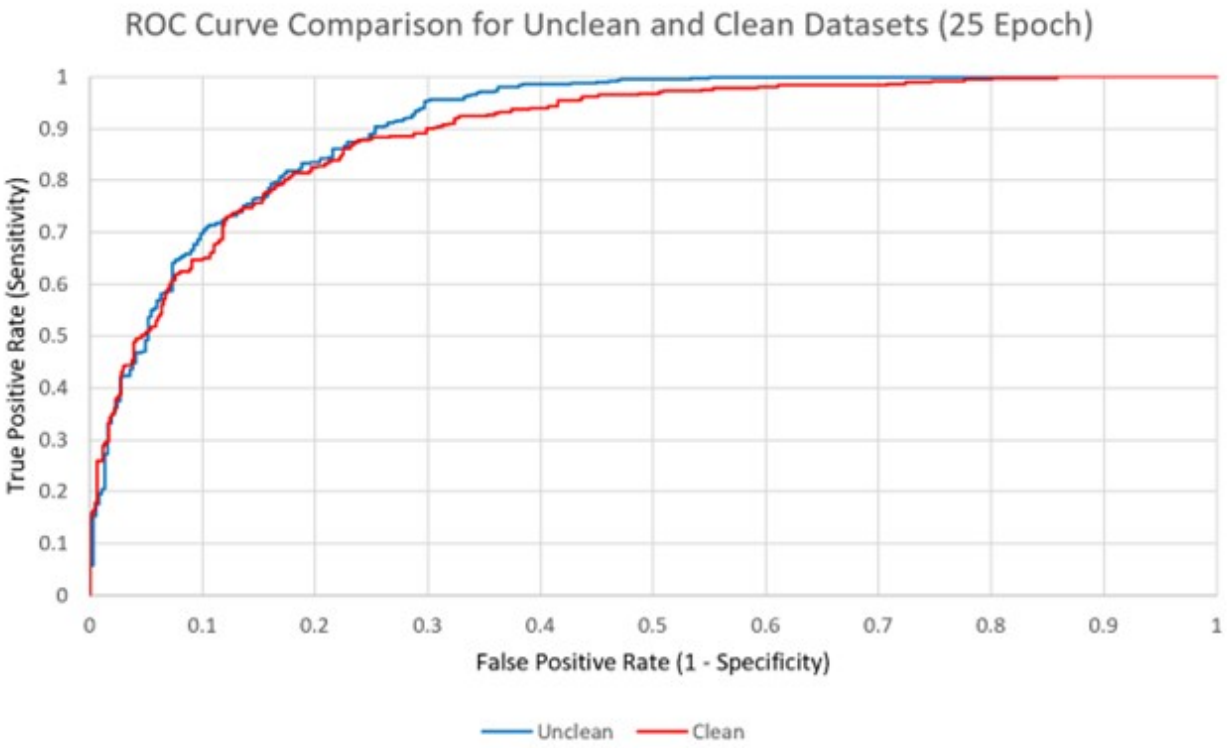

Figure 8: *ROC curves for both datasets at 25 epochs. Both datasets follow a very similar pattern, with the unclean dataset slightly outperforming with a difference of approximately 0.15 AUC. A curve close to the top-left portion of the graph indicates improved performance over all thresholds.*

4.3 Average Precision

Recall is the proportion of total negative predictions that are correct, whereas precision is the proportion of total positive predictions that are correct. Average precision (AP) measures the precision over each recall value and is useful in considering both values simultaneously (Zhu 2004). It can be found by finding the area under the graph of precision against recall. Figure 9 displays how the precision varies with recall in both datasets at 25 epochs. The unclean dataset has a noticeably higher AP, at 0.9378, compared to the clean dataset, with an AP of 0.8384, suggesting that the model produced fewer false results on average when trained on the unclean dataset.

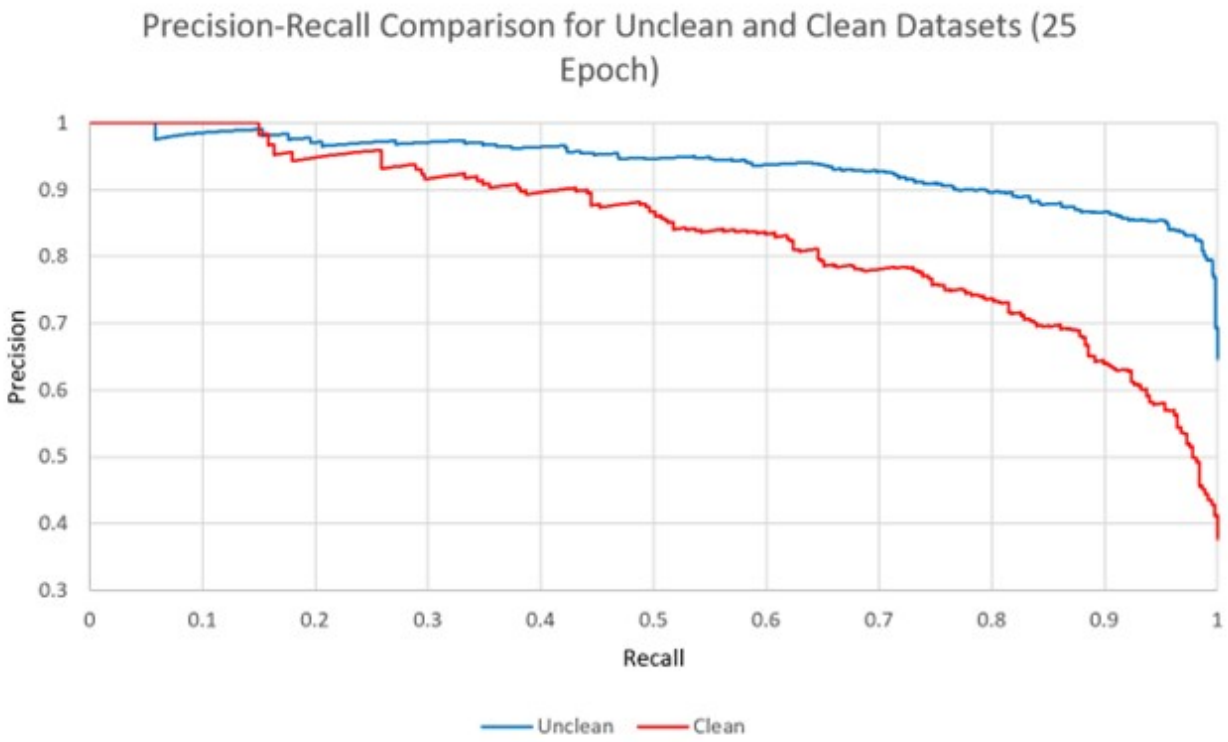

Figure 9: *Precision-recall curves for both datasets at 25 epochs. For the clean dataset, the precision drops at a faster rate at almost all recall values, compared to the unclean dataset which steadily decreases, hence the clean dataset having a lower AP value.*

4.4 Classification Threshold

The classification threshold indicates how strict the model classifies a lesion as malignant. The higher the threshold, the higher the specificity and lower the sensitivity, because as the model becomes stricter on positive outcomes, it will tend towards producing more negative outcomes, and vice versa. An ideal threshold should be strict enough to avoid false positive outcomes, but not too lenient where every lesion is classified as benign. Figure 10 shows graphs of the sensitivity, specificity, and F1 score, which is the combination of precision and recall, against the classification threshold at 25 epochs for both datasets. As expected, the sensitivity decreases and specificity increases as the threshold increases in both graphs. However, the F1 score in the unclean dataset increases and reaches a maximum of approximately 0.92 at a threshold value of 0.5, but the clean dataset has an F1 score of approximately 0.63 at 0.1 classification threshold and subsequently decreases at higher thresholds, indicating higher model stability on the unclean dataset.

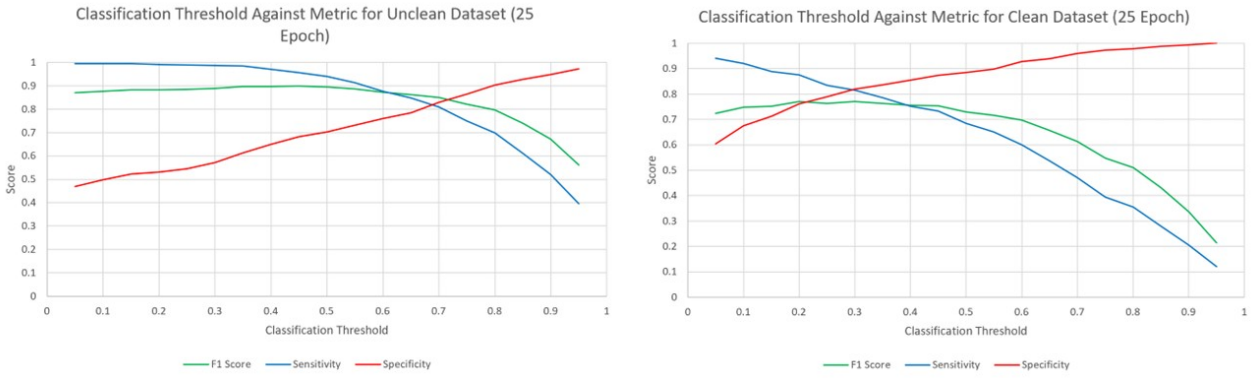

Figure 10: *Classification threshold against sensitivity, specificity, and F1 score for both datasets at 25 epochs. The sensitivity and F1 score drops at a significantly faster rate with the clean dataset, whilst the specificities in both datasets increase at similar rates due to the model shifting towards negative outcomes.*

4.5 Confusion Matrices

Finally, confusion matrices were created to represent how sensitivity, specificity and overall accuracy differed between the two datasets. Figure 11 shows the matrices resulting from the model's predictions at 25 epochs. The unclean dataset resulted in a higher number of true positives, whilst the clean dataset showed a higher proportion of true negative outcomes. Like previous metrics, these matrices indicate that the model performed at a higher accuracy of 84.8% on the unclean dataset, compared to 79.4% for the clean dataset, which is likely due to the significant decrease in sensitivity from the unclean to the clean dataset, whilst the specificities remained relatively similar.

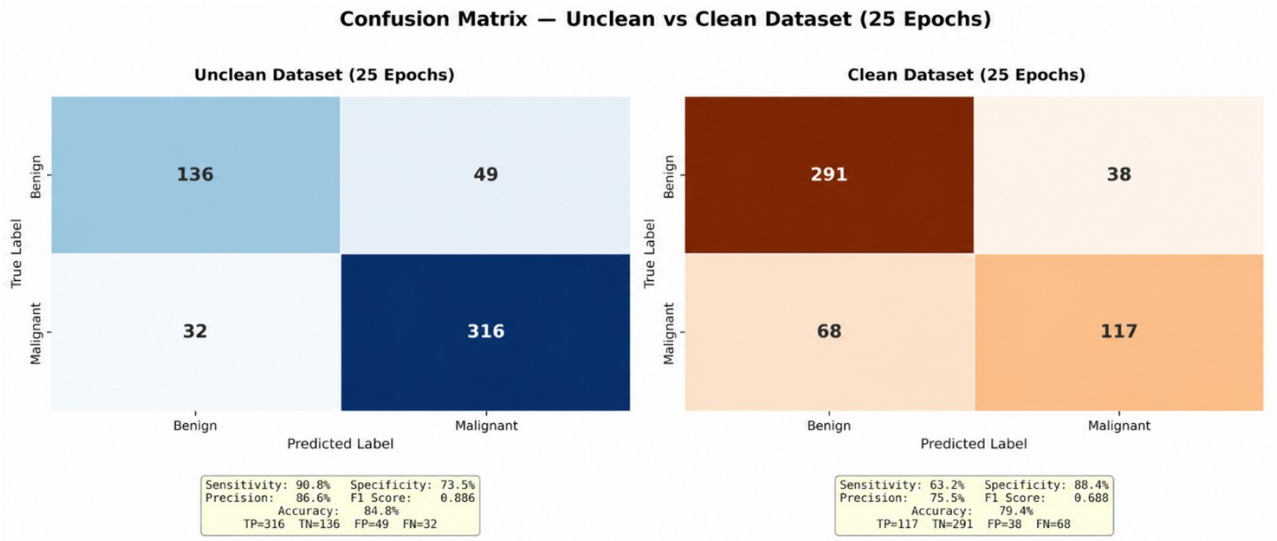

Figure 11: *Images showing (a) confusion matrix for the unclean dataset at 25 epochs, and (b) confusion matrix for the clean dataset at 25 epochs. Overall, the sensitivity was significantly higher for the unclean dataset compared to the clean dataset. Specificity results show that the model detected a higher proportion of true negatives in the clean dataset. However, the overall accuracy was higher for the unclean dataset due to the significant decrease in sensitivity of the model on the clean dataset.*

4.6 Multi-Classification Model

The multi-classification model was also attempted; however, there was limited time to train and validate the model sufficiently, leading to results that did not achieve a suitable standard to include in the main body of this report. Furthermore, it was decided that the mobile application would include the binary classification model due to a higher level of training, validation and testing.

5 Analysis and Findings

5.1 Epoch & Performance Metrics

These results give an understanding of the effect of changing the epoch value on various performance metrics of a machine learning classification model. As expected, the higher epoch values of 20 and 25 performed with a greater accuracy for both clean and unclean datasets. This can be attributed to increased optimisation of parameters such as colour or edge identification because of a longer training duration, as well as having higher exposure to the data with more passes, resulting in increased generalisation. Therefore, epoch value was directly proportional to accuracy up to a limit which would cause overfitting.

With regards to the specificity values, 15 epochs and 20 epochs performed the best in the unclean and clean datasets, respectively. The 15-epoch result may have been due to the model understanding general relationships in the unclean data, whilst not overfitting to noise, hence increasing true negatives and decreasing false positives.

On the clean dataset, 20 epochs allowed a longer training duration, letting the model identify subtle features and maintain consistency across benign results, hence improving the specificity. At 25 epochs, the model may have overfitted in the unclean dataset, and in the clean dataset, slightly overfitted towards the clean images, resulting in the model recognising features irrelevant to the diagnosis of lesions. Finally, as predicted, the results for the 10-epoch value shows the poorest performance, likely due to a low training duration.

The results for sensitivity show that an epoch value of 10 was optimal for both datasets. However, this is almost certainly because of underfitting, as the model may have been biased towards positive outcomes. This is shown when looking at the true and false positives, as the model with 10 epochs identified the highest number of true positives, but also the highest number of false positives compared to the other epoch values. This is further reinforced by looking at the specificity for the 10-epoch results, performing the worst in identifying true negatives and therefore had the lowest specificity in both datasets, suggesting that the model was skewed towards positive results, and the sensitivity value for 10 epochs was anomalous. The sensitivity for the other epoch values followed an expected pattern, where a higher number of epochs (up to a limit) lead to a higher sensitivity, as with more runs, the model is gradually optimised to identify true positives while avoiding false positives.

Similar to the specificity, 15 and 20 epochs are optimal for precision in the unclean and clean datasets, respectively. At 15 epochs, the model does not overfit to noise in the unclean data, basing the diagnosis on obvious malignant features, and at 20 epochs, the model has a longer training duration, and as the clean dataset has reduced artefacts, the model can notice subtle yet relevant features to include in the diagnosis. At 25 epochs, the precision decreases very slightly in both cases, which could result from overfitting. However, the change is relatively insignificant, indicating that an epoch value between 20 and 25 for the clean dataset, and between 15 and 25 for the unclean dataset would be optimal. An epoch value of 10 presents the lowest precision, which can be explained by the skewness of the model towards malignancy, resulting in a higher number of false positives.

The final metric to compare is the F1 score for each epoch value. At 25 epochs, the model showed the highest F1 score for both datasets, which was anticipated as the longer training duration would theoretically lead to higher precision and recall, hence reducing the overall percentage of false results, and increasing the percentage of correct positive cases. An unusual result is seen when comparing the F1 scores for 10 and 15 epochs, which is higher for 10 epochs. After analysing the other metrics, it was determined that this is due to the model's bias towards positive outcomes. Therefore, as the recall value at 10 epochs is much higher than 15 epochs, when weighing out precision and recall in the F1 score, it shows a greater value at 10 epochs.

5.2 ROC & Precision Recall Curves

From the ROC and AUC values, the unclean dataset slightly outperforms the clean dataset. When the false positive rate was 0 to 0.25, the true positive rate of both datasets increased proportionally at the same rate, indicating that the model was being trained consistently at lower thresholds. From a false positive rate of 0.25 onwards, the sensitivity of the clean dataset tapers off, whilst the unclean dataset tapers off at a false positive rate of approximately 0.35. The plateaus are likely a result of the threshold value; as the model becomes stricter on classifying malignancy, the true positive rate decreases. However, as the clean dataset plateaus sooner, the model may have been producing negative outcomes at a faster rate, hence reducing true positive rate.

The precision-recall curves and AP values show that the unclean dataset produced a notably reduced number of false results. This could have potentially arisen from the model overfitting to the clean images. As a result, the model may have produced criteria for malignancy based on unrelated features from the lesion images, hence increasing the number of false positive cases and reducing the precision. The unclean dataset maintains a steady decrease in precision over time at each recall value, whereas the clean dataset caused the model's precision to decrease at a higher rate, ending with a lower precision at the maximum recall value, suggesting that the model was more stable when trained on the unclean dataset.

5.3 Unclean vs. Clean Dataset Comparison

The purpose of this activity was to investigate the impact of curating a clean dataset on machine learning model performance. As seen from these results, at higher epochs, the model may overfit to noise in an unclean dataset, resulting in a diagnosis which is based on some irrelevant features of each image, typically leading to a higher number of false outcomes. However, the results showed a similar number of false positives for both datasets, and unexpectedly, a significantly higher number of false negatives on average for the clean dataset. A cleaner dataset usually results in labels with higher quality and consistency than an unclean dataset, which can increase clarity in benign features. However, this simultaneously narrows the definition of a malignant lesion, where the model develops stricter requirements to produce a malignant outcome. Therefore, this may explain the higher number of false negatives obtained in the clean dataset. This implies that the model was not exposed to enough borderline malignant cases and is less representative of the variability in real-life data.

6 Discussion, Evaluation & Implications

This report studied the performance of a machine learning classification model in detecting malignancy in skin lesions, which was then integrated into a mobile application for GP use. The model was tested on two types of datasets – an unclean dataset, which involved picking out images from renowned public databases, and then a clean dataset where more scrutiny on artefact removal was applied, hence removing and replacing some images in a cleaning process, with the sub-aim of testing the impact of quality control in dataset curation on model performance. Also, the effect of adjusting the epoch value, a model parameter which controls the number of passes the model makes across the dataset, on various performance metrics was also investigated. Contrary to what was hypothesised, the unclean dataset slightly outperformed the clean dataset in terms of sensitivity and specificity, as well as precision and accuracy. This was likely due to the clean dataset providing images that were overly cleaned, meaning that the model may have been trained on benign and malignant cases that were very apparent. As a result, the model may have failed to understand characteristics of borderline malignant cases, leading to an increase in false negatives and hampering the performance. This implies the importance of curating a dataset that avoids major artefacts but also provides a suitable representative of real-world clinical data, where significant variations can occur that models must account for. The unclean dataset was not necessarily unclean as suggested, but potentially represented the fluctuations associated with borderline cases in practice better than the clean dataset, hence producing better results. Therefore, curating a dataset with sufficient variability is crucial for model performance, being strict enough to avoid major atrefacts.

To further evaluate the performance of the model, it was compared to other CNN architectures, including EfficientNetV2L, DenseNet201 and ResNet50, the results differed quite significantly. Also, the purpose of using additional models was to test the original hypothesis that the clean dataset should allow for greater performance. In terms of accuracy, AUC of an ROC curve, and precision, the DenseNet201 performed better on the clean dataset with higher values of each metric than the other models, including the one used in this project. Additionally, this CNN resulted in higher performance metrics on the clean dataset as opposed to the unclean dataset, likely due to stabilising because of little variations in image quality. The optimal epoch value for DenseNet201 was determined at around 17-18, which implied less computational expense than this project's model. However, the metrics obtained from the additional three CNNs were not as high as the model trained on the unclean dataset in this project, which suggests that the CNN model in this project was successful. The graphs for accuracy, AUC, precision and loss for the three additional CNNs tested can be found in Appendices.

The ambiguity in these results reinforce the addition of XAI into this project. Much of the Analysis and Findings section is based on assumptions of the model and results from previous studies, meaning that more iterations could have further evidenced the implications of the results in this project. By implementing XAI, the app provided justifications as to which features of a specific lesion was flagged as suspicious, how confident the model is in its outcome of a lesion image, and how the outputs and justifications change based on adjusting the

input data and parameters in the classification model, all providing a clearer understanding of how the model learns to interpret skin lesion images. Furthermore, with the use of GradCAM, a heatmap was generated for each image being analysed, clearly highlighting regions of suspicion. Performance tracking in real-world use is critical to avoid the output data drifting over time; this is a common issue in machine learning models, where over a long period, the input data changes significantly, hence the model adjusting the relationships and criteria that are attributed to benign or malignant lesions (Nelson et al. 2015).

Finally, this model links strongly to the front-end portion of the application, as the app facilitates uploading of images and the subsequent outcome. Therefore, a key requirement of the application that was utilised was to only allow dermoscopic images to be uploaded, which can be done via uploading an image downloaded from the internet, or taken with a dermatoscope. Consequently, the results would have a reduced possibility of being skewed due to completely irrelevant images, allowing the model to perform consistently over many uses, which can be incorporated into GP settings to reduce referrals. Figure 12 shows an example of the model output when a malignant lesion was uploaded to the app, showing XAI output and a heatmap generated by GradCAM.

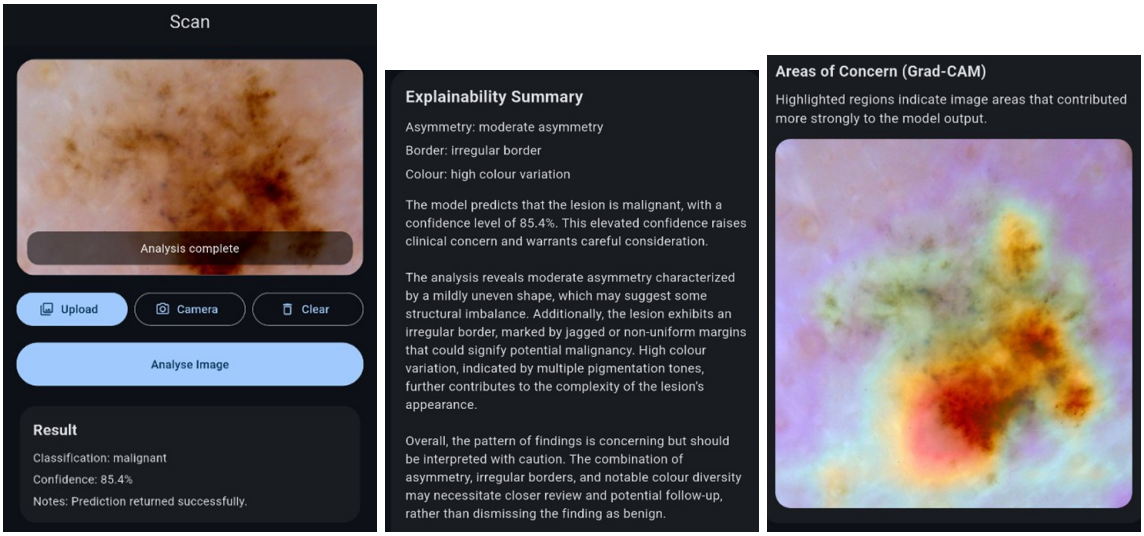

Figure 12: *Images showing (a) image uploading interface and lesion used in the example, (b) the XAI output, justifying the classification and confidence score using the ABCDE criteria, and (c) the heatmap generated for the lesion using GradCAM to highlight suspicious areas.*

7 Future Work

Given more time and resources, the experimental stage of model optimisation could have put more focus on continuously adjusting parameters other than the epoch. For example, by changing the batch size, a more efficient balance of accuracy and computational efficiency could have been achieved, or by refining the dropout value, it could have prevented signs of overfitting that were seen during validation. The combination of these adjustments could have improved the accuracy as well as other metrics, whilst keeping computation time minimal. Additionally, other deep learning architectures mentioned in the Literature Review could have been tested in this application. Although CNNs generally outperform SVMs and KNNs in the literature, it would be beneficial to confirm this on this project's dataset. Finally, more iterations of the multi-class classification model, which was an expected outcome of the project, could have been performed to run a successful model, splitting benign lesions into their respective subtypes. Due to the number of artefactual images which were removed, the dataset was not large enough to sufficiently train, validate and test the multi-class model.

Regarding dataset curation, the process of collecting images ran somewhat efficiently. However, in hindsight, private databases could have been contacted to obtain access to additional images, which could have potentially reduced the number of artefactual images, saving time on filtering through unwanted images. Selecting images

from the ISIC archive and HAM10000 datasets worked well as all images were stored in one space, but the quality of images varied significantly, whereas private databases may have had more stringency on collecting clearer images. Although it was difficult to locate and contact international databases, these would have been useful in collecting an even distribution of images across all FSTs. With more time to curate images, this would be a key protocol in dataset curation.

Gantt Chart

AI and Sensor Development for Skin Cancer Detection — Project Members : Jatinder, Gurtej, Kennasa, Holly, Nadia, Emily

MAIN TASKS	SUB-TASKS	START DATE	END DATE
Dataset Curation and Model Development	Self-analyse of skin lesion images	20-Oct-25	27-Oct-25
	Initial dataset of 200 Images (Cherry Angioma and Intradermal Nevus)	11-Nov-25	24-Nov-25
	Test initial model script given by Brandon on initial dataset	02-Dec-25	15-Dec-25
	Interpret and upload results to shared folder	13-Jan-26	13-Jan-26
	Larger-Scale Dataset Curation (3000 benign and 3000 malignant images)	03-Feb-26	11-Mar-26
	Split benign images into four specific subtypes (750 images each)	11-Mar-26	18-Mar-26
	Split dataset into 70% training, 20% validation, and 10% test data	18-Mar-26	20-Mar-26
	Write up a report on challenges faced when collecting skin lesions	20-Mar-26	23-Mar-26
	Create CNN model using assistance from LLMs, establish parameters	24-Mar-26	01-Apr-26
	Remove artefactual images to create a clean and unclean dataset	30-Mar-26	07-Apr-26
	Test the model on the clean and unclean datasets, recording results in graphs for analysis	07-Apr-26	08-Apr-26
	Refine model parameters to improve performance metrics	09-Apr-26	10-Apr-26
	Write up a report on justifying parameters and evaluating performance of model on both datasets	11-Apr-26	11-Apr-26
Other Workstreams	Image Processing - hair removal, segmentation techniques, U-Net model implementation	20-Feb-26	02-Apr-26
	Fairness/Ethical Monitoring - monitoring bias in datasets across skin tones	03-Feb-26	11-Mar-26
	Fairness/Ethical Monitoring - monitoring bias in model performance across skin tones	07-Apr-26	13-Apr-26
	XAI Output - provide justifications of outcomes with a confidence score + GradCAM heatmap	15-Mar-26	20-Apr-26
	UX and UI Development - studying user requirements, designing a simple but efficient interface	20-Mar-26	20-Apr-26
	Validation/Performance Tracking - monitor app in real-time, minimising model drift over time.	20-Apr-26	28-Apr-26

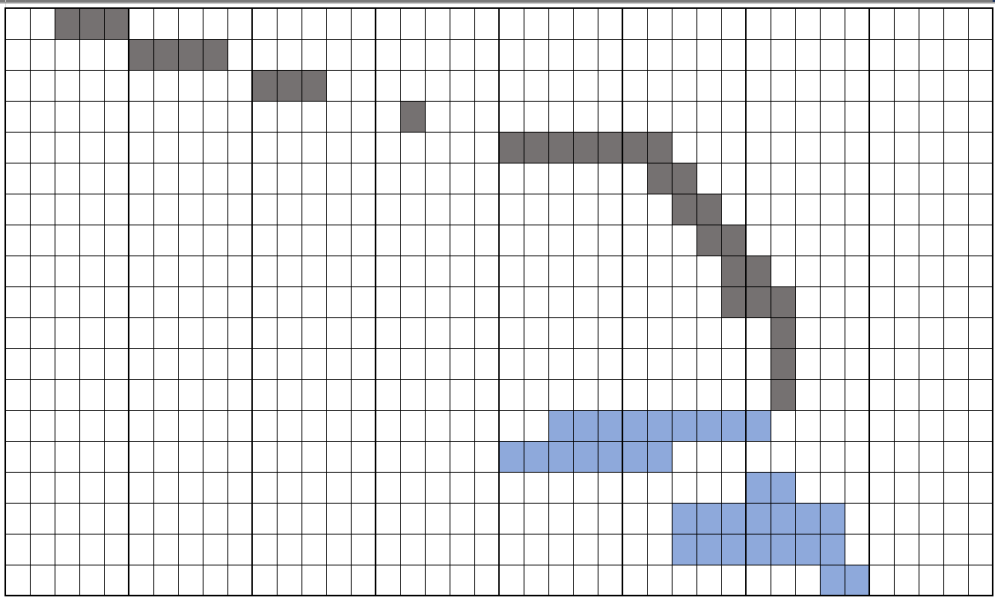

Figure 13: *Gantt chart showing project plan over the course of the year. In grey, the tasks relevant to the individual contribution in this report are listed, and in blue, the additional key tasks that were required to complete the mobile application.*

References

Ascott, A., van Bodegraven, B., Ahmed, S., Harwood, C., Proby, C., Sommerlad, M., Vernon, S., Welsh, C. & Venables, Z. (2023), 'P49 epidemiology of skin cancer by ethnicity in England, 2013–2020: a cohort study', *British Journal of Dermatology* **188**(Supplement_4).

Bashir, D., Montañez, G., Sehra, S., Segura, P. & Lauw, J. (2020), An information-theoretic perspective on overfitting and underfitting, *in* 'AI 2020: Advances in Artificial Intelligence', pp. 347–358.

Bjorck, J., Gomes, C., Selman, B. & Weinberger, K. (2018), Understanding batch normalization, *in* 'Advances in Neural Information Processing Systems'.
URL: *https://proceedings.neurips.cc/paper_files/paper/2018/file/36072923bfc3cf47745d704feb489480 – Paper.pdf*

Cervantes, J., Lamont, F., Mazahua, L. & Lopez, A. (2020), 'A comprehensive survey on support vector machine classification: Applications, challenges and trends', *Neurocomputing* **408**(1), 189–215.

Daneshjou, R., Vodrahalli, K., Novoa, R., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S., Bailey, E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J., Okata-Karigane, U., Zou, J. & Chiou, A. (2022), 'Disparities in dermatology AI performance on a diverse, curated clinical image set', *Science Advances* **8**(32).

Edge Health (2024), 'NHSE outpatient recovery and transformation programme: Evaluating pathways for AI dermatology in skin cancer detection', https://www.edgehealth.co.uk/wp-content/uploads/2024/08/Evaluating-Pathways-for-AI-Dermatology-in-Skin-Cancer-Detection.pdf.

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H. & Thrun, S. (2017), 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature* **542**(7639), 115–118.

Goddard, K., Roudsari, A. & Wyatt, J. (2012), 'Automation bias: a systematic review of frequency, effect mediators, and mitigators', *Journal of the American Medical Informatics Association* **19**(1), 121–127.

Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A., Thomas, L., Enk, A., Uhlmann, L. & et al. (2018), 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists', *Annals of Oncology* **29**(8), 1836–1842.

Hu, L.-Y., Huang, M.-W., Ke, S.-W. & Tsai, C.-F. (2016), 'The distance function effect on k-nearest neighbor classification for medical datasets', *SpringerPlus* **5**(1).

Hussein, B. & Shareef, S. (2024), 'An empirical study on the correlation between early stopping patience and epochs in deep learning', *ITM Web of Conferences* **64**, 01003.

Indolia, S., Goswami, A., Mishra, S. & Asopa, P. (2018), 'Conceptual understanding of convolutional neural network — a deep learning approach', *Procedia Computer Science* **132**, 679–688.

ISIC Archive (n.d.), 'ISIC archive', https://gallery.isic-archive.com/#.

Kandel, I. & Castelli, M. (2020), 'The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset', *ICT Express* **6**(4).

Kataria, A. & Singh, M. (2013), 'A review of data classification using k-nearest neighbour algorithm', *International Journal of Emerging Technology and Advanced Engineering* **3**(6), 354–360.

Levell, N. (2021), 'Dermatology GIRFT programme national specialty report', https://gettingitrightfirsttime.co.uk/wp-content/uploads/2021/09/DermatologyReport-Sept21o.pdf.

Liu, P., Qian, W. & Wang, Y. (2024), 'YWnet: A convolutional block attention-based fusion deep learning method for complex underwater small target detection', *Ecological Informatics* **79**, 102401.

MASCED (2025), 'The ABCDE of melanoma method', https://masced.uk/abcde-of-melanoma.htm.

MHRA (2019), 'Medical device stand-alone software including apps (including IVDMDs)', UK Government Policy.

Nahm, F. (2022), 'Receiver operating characteristic curve: overview and practical use for clinicians', *Korean Journal of Anesthesiology* **75**(1), 25–36.

Nelson, K., Corbin, G., Anania, M., Kovacs, M., Tobias, J. & Blowers, M. (2015), Evaluating model drift in machine learning algorithms, *in* 'Proceedings of the IEEE Conference on Computational Intelligence and Data Science and Analytics (CISDA)'.
URL: *https://doi.org/10.1109/CISDA.2015.7208643*

NICE (2022), 'Digital technologies for the detection of melanoma: Medtech innovation briefing', https://www.nice.org.uk/advice/mib311/resources/digital-technologies-for-the-detection-of-melanoma-pdf-2285967623291845.

NICE (2025), 'Artificial intelligence (AI) technologies for assessing and triaging skin lesions referred to the urgent suspected skin cancer pathway: early value assessment', https://www.nice.org.uk/guidance/hte24.

Öztürk, Ş., Özkaya, U., Akdemir, B. & Seyfi, L. (2018), Convolution kernel size effect on convolutional neural network in histopathological image processing applications, *in* 'Proceedings of ISFEE 2018'.

Patient.info (2022), 'Benign skin tumours', https://patient.info/doctor/dermatology/benign-skin-tumours. Accessed: 20 April 2026.

PureLite (2021), 'Fitzpatrick scale — skin types', https://www.purelitekent.co.uk/fitzpatrick-scale-skin-types/.

Rukundo, O. (2023), 'Effects of image size on deep learning', *Electronics* **12**(4), 985.

Skene, S., Moschoyiannis, S., Armes, J., Sutton, K., Sutton, S., Quentin, L., Demestihas, M.-A., Jenkins, H., Mendis, J. & Unity Insights (2024), 'AI in healthcare award — final report: Skin analytics (DERM)', Surrey Open Research Repository (University of Surrey).

Skin Analytics (2023), 'Regulatory certification', https://skin-analytics.com/news-category/regulatory-certification/.

Skinive (2022), 'Skinive accuracy report 2022', https://skinive.com/skinive-accuracy2022/. Accessed: 19 April 2026.

SkinVision (2025), 'Pricing and refunds', https://skinvision.zendesk.com/hc/en-gb/articles/22945140402844-Pricing-and-Refunds.

Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. (2014), 'Dropout: A simple way to prevent neural networks from overfitting', *Journal of Machine Learning Research* **15**, 1929–1958.
URL: *https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf*

Tsao, H., Olazagasti, J., Cordoro, K., Brewer, J., Taylor, S., Bordeaux, J., Chren, M.-M., Sober, A., Tegeler, C., Bhushan, R. & Begolka, W. (2015), 'Early detection of melanoma: Reviewing the ABCDEs', *Journal of the American Academy of Dermatology* **72**(4), 717–723.

Wen, D., Khan, S., Xu, A., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., Perez, C. d. B., Denniston, A., Liu, X. & Matin, R. (2021), 'Characteristics of publicly available skin cancer image datasets: a systematic review', *The Lancet Digital Health* **4**(1).

Zhu, M. (2004), 'Recall, precision and average precision', https://datascience-intro.github.io/1MS041-2022/Files/AveragePrecision.pdf.

A Tables

Table A1: *Performance metrics for the unclean dataset evaluated at five-epoch intervals.*

	Epoch			
Metric	**10**	**15**	**20**	**25**
Sensitivity (%)	95.98	78.45	90.80	90.80
Specificity (%)	54.59	78.92	68.65	73.51
Accuracy (%)	81.61	78.61	83.11	84.40
Precision (%)	79.90	87.50	84.49	86.58
F1 Score	0.8721	0.8273	0.8753	0.8864
True Negatives	101	146	127	136
False Positives	84	39	58	49
False Negatives	14	75	32	32
True Positives	334	273	316	316

Table A2: *Performance metrics for the clean dataset evaluated at five-epoch intervals.*

	Epoch			
Metric	**10**	**15**	**20**	**25**
Sensitivity (%)	69.73	58.38	61.08	63.24
Specificity (%)	80.85	87.84	90.58	88.45
Accuracy (%)	76.85	77.24	79.96	79.38
Precision (%)	67.19	72.97	78.47	75.48
F1 Score	0.6844	0.6486	0.6869	0.6882
True Negatives	266	289	298	291
False Positives	63	40	31	38
False Negatives	56	77	72	68
True Positives	129	108	113	117

B Figures

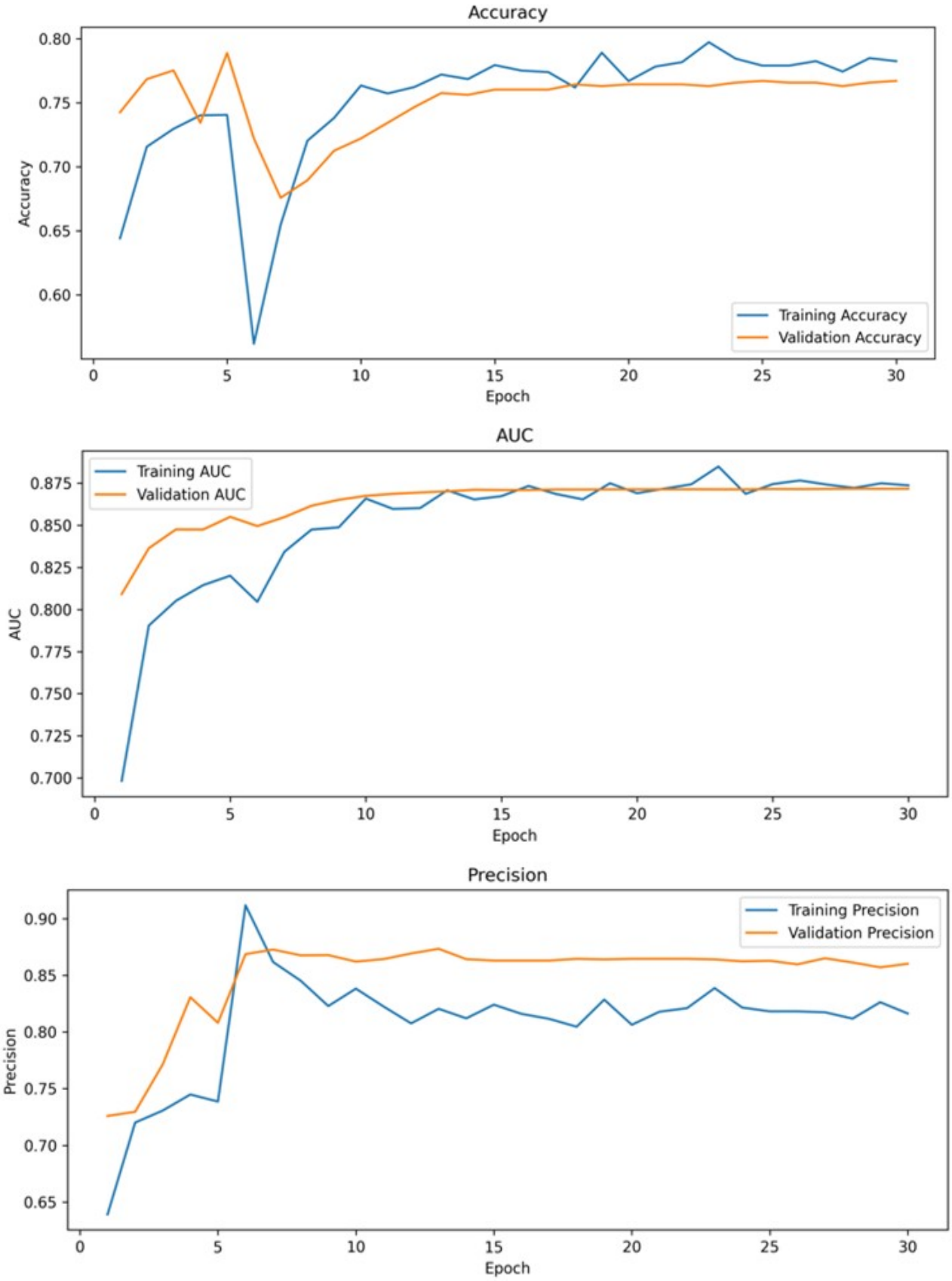

Figure B1: *Images showing (a) accuracy against epoch, (b) AUC against epoch, and (c) precision against epoch for DenseNet201 CNN architecture.*