APPLIED MACHINE LEARNING SYSTEMS II (ELEC0135) 24/25 REPORT: MELANOMA CLASSIFICATION FOR SIIM-ISIC KAGGLE COMPETITION

SN: 24251714

ABSTRACT

This research explores resource-efficient approaches to melanoma classification from dermoscopic images, addressing the challenge of operating under computational constraints in the context of the SIIM-ISIC 2020 competition. While top-performing submissions typically employ ensemble methods with multiple models trained on extensive datasets over several days using powerful GPU clusters, we demonstrate that competitive performance can be achieved with significantly fewer resources through strategic data sampling and model selection. Our approach focuses on constructing a smaller, high-quality dataset of 3,987 images (compared to the original 33,126) by maintaining diversity across patients while preserving all melanoma cases, which resulted in a model that achieved a ROC-AUC of 0.845 on the competition test data. This represents a substantial efficiency improvement, reducing training time from hours to minutes while maintaining performance comparable to models trained on the full dataset. We also investigated specialized models based on sex and image brightness characteristics, though these underperformed due to insufficient data for separate feature extractors. Our findings suggest that in medical imaging applications with highly imbalanced datasets, thoughtful dataset construction can be more valuable than raw data volume, making accurate melanoma detection more accessible in resource-limited healthcare environments. This work contributes to the broader effort of developing efficient AI solutions for medical applications where computational resources may be limited but diagnostic accuracy remains critical. [1].

1. INTRODUCTION

Machine learning is widely recognized as a transformative force in the medical industry, with one of its most promising applications being medical imaging—particularly image classification. This technology offers the potential to assist in diagnosing medical images without requiring the constant involvement of highly specialized doctors. As a result, it could significantly reduce pressure on healthcare resources. In the UK, doctors are already facing high levels of demand, and dermatologists are no exception. Dermatology, in particular, stands to benefit greatly from advancements in medical imaging powered by machine learning.

Skin cancer is one of the most common forms of cancer, with melanoma being the fifth most prevalent in the UK—accounting for 5% of all new cancer cases between 2017 and 2019.[1] Melanoma is the most invasive type of skin cancer and carries the highest risk of mortality, making it a key focus in the field of skin cancer image detection.

Traditionally, the diagnostic process for melanoma begins with a visual examination of the mole. Dermatologists assess specific characteristics commonly associated with melanomas before deciding whether to proceed with further steps such as mole removal, biopsy, and treatment.[2] Early detection is critical to improving survival rates, which makes widespread initial screening essential. Enabling efficient and accessible early-stage assessments for as many patients as possible is therefore a top priority. [3]

The Society for Imaging Informatics in Medicine in collaboration with the International Skin Imaging Collaboration (SIIM-ISIC) often run image classification challenges to encourage research into the area. There have been several competitions over the years with the most recent in 2024 comprising of a competition for skin cancer detection with 3D total body scans. The 2020 competition [2]was chosen as it was more accessible to those with limited resources. The competition involves classifying medical images of moles as malignant or benign with accompanying tabular data including age, sex and location to potentially assist with classification performance. The performance of each of the submissions was based on their Receiver Operator Curve Area Under Curve (ROC-AUC) score. All of the top submissions on Kaggle required extensive amounts of compute to achieve competitive performance. Therefore, the purpose for the research in this report was to explore a novel approach to completing this competition in a resource-constrained environment by better exploring the underlying dataset.

2. LITERATURE SURVEY AND BACKGROUND

Kaggle competitions allow participants to discuss and share solutions to the competitions. This can be enormously beneficial for research purposes but can often be detrimental in a

[1]The code for this paper can be found here.

[2]The kaggle competition can be found here.

competition setting. Often what happens is individuals copy solutions from other high performing individuals and tweak some hyperparameters or run a slightly more complex model to achieve marginal improvement in the model performance. This leads to huge homogeneity in the leader board with little to no additional thought about what specifically could be done to improve the performance of the model. Additionally, competitors often use significant external compute running on clusters of commercial grade GPUs for several days. Although this often is one way to drastically improve model performance as more data can be used to train the model, it is also less accessible for those without the compute available or for applications in a resource constrained environment.

Additionally, almost all of the submissions for this competition made use of data that was released from previous years and combined it with the data from the 2020 competition. This was done due to the highly imbalanced nature of the dataset, containing only 1.76% positive samples. The primary affect this had on the dataset was to stabilize the model performance. Although effective, it technically compromises the underlying purpose of the challenge which is to apply classification to extremely imbalanced data.

The winner of the Kaggle competition [4] made use of an ensemble of 18 image classifiers to achieve an average ROC-AUC of 0.9442 on the test dataset. Ensemble methods work by combining the outputs for several models into a singular prediction which generally performs better than any singular model. It is a very common approach and was used extensively by other competition participants, by varying the image sizes and underlying models being trained.

The issue with these models is that they often take significantly longer to train with limited resource. Although impressive, the 18 separate models combined with increased number of images, larger image sizes and 5-fold cross-validation meant that it took several days to train on 8 Nvidia Tesla V100 GPUs. For most, this would require some form of cloud compute for training. Despite GPU cloud computing becoming more available and affordable in recent years, it may be desirable to hold onto data and only perform computing offline in remote environments or for data security purposes.[5] The authors of the paper only use the tabular data on 4 of their 18 models, suggesting that they deemed it less important for predictions.

This hypothesis that additional information does not affect model performance somewhat aligns with research conducted by K. Sies et al.[6] into sex bias of CNNs applied to melanoma classification. The researchers concluded that a market-approved CNN for skin cancer classification (MoleAnalyzer Pro, Fotofinder Systems GmbH, Bad Birnbach, Germany) did not have sex related biases in its ability to carry out melanoma classification. This is despite sex-related imbalances in the ISIC database on 11th October 2021 which had 39.2% of all images classified as female, 48.0% as male and 12.9% as unknown. It should be noted that in this context sex is defined as the biological definition of an X or Y in the 23rd chromosome, and this definition is maintained throughout this paper.

The image classifiers forming the backbone of the majority of the models submitted on Kaggle were mostly some variety of EfficientNet [7] or ResNet [8]. Both of these models are forms of convolutional neural networks also known as CNNs. EfficientNet is a special model in that it actively adapts the depth and width of the model based on the input image size making it highly efficient for a range of uses. ResNet-18 implements an additional feature called residual connections which act to reduce the effect of a property deep learning networks called the vanishing gradient problem. It works essentially by adding the weights from previous layers to the current layer, i.e.$\mathbf{x} + \mathcal{F}(\mathbf{x}) =$. A visualization of this is seen below in figure 1.

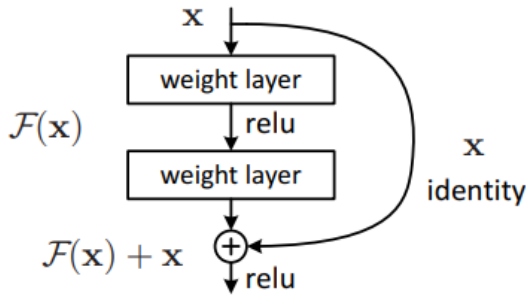


Fig. 1. Diagram taken from original ResNet Paper showing the functioning of a residual block in the CNN. [8]

There has been significant research into using AI models for Melanoma classification with products now available on the market for this exact task. It has been shown that dermatologists and medical practitioners formally trained in identifying melanomas only had an average sensitivity of ¡80%. [9] In 2018, H.A. Haenssle et al. trained an off-the-shelf Google Inception v4 CNN architecture using 100,000 images of dermatoscopic images. [10] The study compared the model's ability to classify melanomas against 58 dermatologists, including 30 experts, and showed that most dermatologists were outperformed by the model. It demonstrates the importance in the technology for all practitioners both experienced and novice.

However, there has been some pushback in the medical community to using AI. Producing models that are both interpretable and explainable has been shown to be an important factor in building trust and transparency in AI [11, 12] and some medical applications give this as a priority over raw performance [13]. Efforts to produce explainable ML models can be split into two classes: inherent and post-hoc.[14] Models such as decision trees are directly explainable and so have been more widely adopted by the medical community.

[15, 16] This is in contrast to Deep Learning models which are often considered black boxes. Several post-hoc methods attempt to explain the outputs of these models by traversing the neural network and often producing a heat map overlaid onto the input image.[17, 18] However, these methods are often criticized for various reasons, including lack of reliability and introducing biases by humans over trusting computer systems particularly those that are complex and difficult to interpret.[14, 15, 16]

3. DESCRIPTION OF DATA

The competition includes 33,126 training samples with 10,982 additional for testing with no label given for competition purposes. The images in the dataset are super high resolution dermoscopic images of moles, often 4000x7000 in size, making the complete dataset over 100GB in size.

The dataset also contained four additional data points for sex, approximate age, location of mole on body and diagnosis of mole. It was found that the number of males and females in the dataset were relatively balanced at around 55% for males. The distribution of the age ranges is centred around the 50 year old range in a normal shaped distribution. The location of the moles was primarily located on the torso and the diagnosis was almost entirely in the unknown category providing very little information.

Additionally, the dataset contains a column for patient ID. Some patients had 115 images of moles in the dataset, however the majority of patients had between 0 and 5 samples in the dataset. This aspect was important during the construction of cross-validation and training datasets as described in subsection 5.1. As a result, it was hypothesized that using less data may be possible to achieve good performance by sampling a small number of images from each patient.

As highlighted in section 2 there is a heavy class imbalance in the data prompting many on Kaggle to use additional data sources for training. To explore the effect this class imbalance had on the data categories, several violin plots were made to compare the relationships between the age, sex, and mole location with the number of melanoma samples. A sample of one these plots can be seen in the appendix in Figure 13.

The distribution of the data revealed distinct differences between males and females in both the typical locations and ages at which melanomas occurred. For example, Females had only one recorded case of melanoma on the palms or soles, whereas males had several. In contrast, males had only one occurrence of melanoma on the oral or genital areas, while females had multiple cases in those regions.

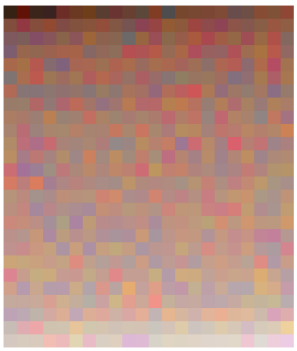
The primary objective of this data exploration was to uncover potential areas where a mixture of experts (MoE) model could offer tangible benefits. Specifically, the aim was to determine whether this modeling approach could help exploit meaningful patterns or subgroup-specific characteristics in the data, thereby enhancing the overall learning process and model performance. MoE models are particularly useful when different subsets of data exhibit unique traits that a single, generalized model may struggle to capture effectively.

As highlighted by K. Sies et al. [6], convolutional neural networks (CNNs) trained on a large dataset of over 100,000 dermoscopic images demonstrated minimal to no bias in sensitivity when predicting melanomas across male and female patients. This finding suggests that, when trained on extensive and diverse datasets, CNNs may be capable of learning representations that generalize fairly across demographic groups. However, this may not hold true in the context of the current study, which relies on a significantly smaller dataset with a much more limited number of positive melanoma cases. In such scenarios, biases are more likely to emerge, and the generalization capabilities of standard models may be impaired, making a strong case for more adaptive or specialized modelling strategies, such as mixtures of experts.

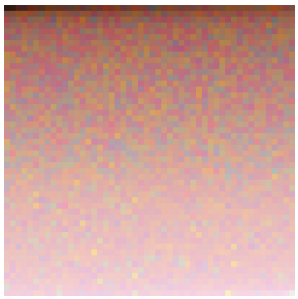
In addition to sample size limitations, their report [6] also emphasized noteworthy differences in the visual features present in male versus female dermoscopic images. For example, it was found that images of male patients were far more likely to contain visible hair, with 22.4% of the male training dataset exhibiting this feature. Such disparities in image characteristics can create challenges for models trained on small, imbalanced datasets, especially when the presence of hair or other artifacts affects lesion visibility or distracts from clinically relevant features. These kinds of image-level variations could introduce unintended biases or reduce model accuracy, particularly if the model fails to appropriately account for subgroup-specific visual cues.

To explore better the portion of data contained in the dataset with hair the images a random sample of 2500 images from the complete dataset along with all melanomas (positive classification) were processed using an averaging filter to obtain the general colour of each image. These images were then sorted by brightness and plotted together in the two matrices shown in Figure 2. The purpose of this was to identify the differences between the colours of images in the melanoma dataset and that of the rest of the images. In the figures it can be seen that the melanoma images are generally far darker in colour than of the rest of the dataset. This observation may have unwanted affects on the model performance, possibly failing to generalize on patients with lighter skin.

To explore this further, two plots were generated, one containing the five darkest and lightest images of the melanomas (figure 3) and one of the benign moles (figure 4). The purpose of this was to understand the nature of the melanomas present in the dataset. In the figures we see that to the untrained eye the lighter images of both models are quite similar. However the darker images of the benign moles are characterized by additional features such as hair or lack of artificial light, whereas for melanomas the darkness appears to the a result of the mole taking up the majority of the frame.



(a) All Melanoma Images Background Colours



(b) 2500 image sample of background colours

Fig. 2. Processed Images Average Colours displayed in matrices.

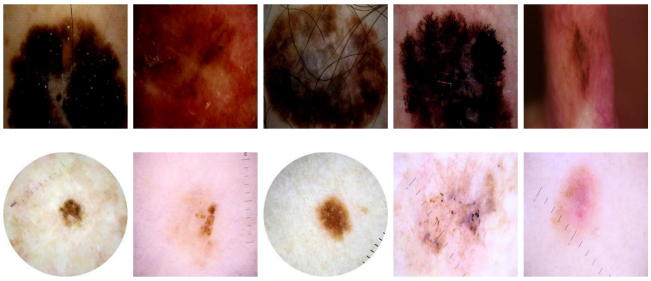


Fig. 3. The five lightest and darkest images of melanomas in the dataset.

This highlighted a potential area to investigate when evaluating model performance. For example, if the best performing model struggles with dark images it may be a result of hair covering the melanoma as opposed to the melanoma itself. It also provides valuable information into model design as training a single classifier on darker images and one on lighter ones may allow it to learn more quickly to ignore the affects of hair rather than learn to classify all darker images a melanomas.

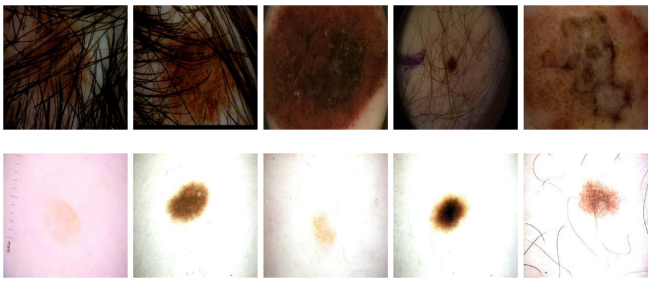


Fig. 4. The five lightest and darkest images of benign moles in the dataset.

4. DESCRIPTION OF MODELS

Convolutional Neural Network deep learning architectures have shown to be some of the most effective architectures for image classification of Melanoma's often outperforming Dermatologist.[19][10]. However, these models often require extensive amounts of data to effectively train as well as large amounts of resources to effectively tune the hyperparameter of the model. These models can be built by hand but often require either deep expertise and hand-on experience to know how many layers to use and what size of layer to use. As a result, for most using a pre-built CNN that has been pretrained on large amounts of data is the preferred choice.

Pre-built models allow us to use architectures that have shown good performance across a range of applications. They also often implement more advanced features that help to negate some of the consequences of using deep learning. Using pre-trained weights allows models to be fine-tuned for specific tasks on smaller amounts of data. This comes from the models learning basic shapes and features on the pre-training dataset which can then be used to extrapolate to other problems. This is what is commonly known as transfer learning, and helps to overcome the problem of overfitting to the data which is where the model fails to learn features that are important in the training data. [20]

CNNs are a form of Neural Network (NN) that have been developed to be particularly well suited to 1D temporal signals such as sensor reading data and 2D signals like images. They leverage local receptive fields within an image to extract information that is tied together to reduce the number of parameters. This is extremely powerful allowing useful features that have been identified in the image to be reused everywhere else without having to be independently learned, resulting in translation invariance.

These local receptive fields act as feature maps which filter the data using the sliding window technique, which convolves some weight matrix across the inputs of the image, adding a bias and then applying a nonlinearity function to the result. This is often done in multiple layers which then feed into a fully connected NN. A simple diagram of a CNN can be seen in Figure 5. The addition of a max pooling layer acts to subsample the output in order to reduce the size and obtain a small amount of shift-invariance and is commonly used in CNNs due to better model performance.[21]

A NN consists of a series of affine transforms followed by some nonlinear activation. Each output and input are said to be fully connected thus resulting in a network of so-called perceptrons. If this nonlinear mapping is a logistic function then the whole model can be thought of as a series of logistic regression functions. An example of a neural network along with a single perceptron has been drawn in Figure 6. Whereby for one perceptron there is the following activation function, where g is the nonlinear transformation, w represents the weights that are learned during training, a is the ac-

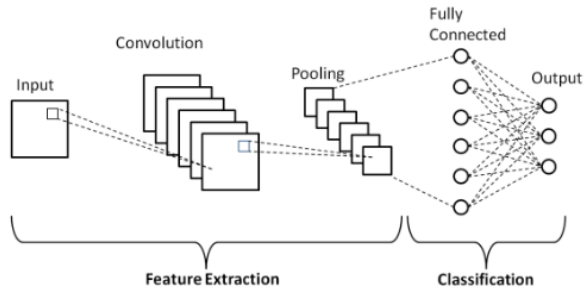


Fig. 5. Schematic of simple CNN, taken from [22].

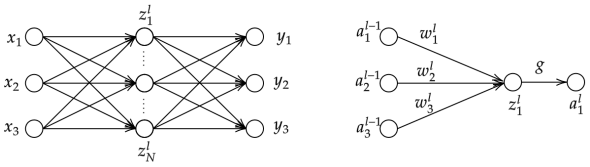


Fig. 6. Left: Simple schematic of a fully connected NN. Right: Schematic of a simple perceptron.

tivation output and b is the bias applied to each input:

$$a_1^l = g\left(\sum_{i=1}^{N} w_N^l \cdot a_N^{l-1} + b_1^l\right) \tag{1}$$

For multiclass classification tasks an additional softmax transformation is applied at the output layer which essentially acts a function to map the relative size of the model outputs to a probability distribution, this is required for NN so that the output can be interpreted as a confidence interval of its predictions. For binary outputs the sigmoid function is used which is essentially the softmax for a single output.

Traditionally, the parameters of CNNs are learned via a gradient-based optimization method that aims to minimize the error of some loss function that defines the relationship between predictions and the outputs for supervised learning tasks. To compute the weights of the model efficiently requires efficient computation of the gradient of the loss function. This is achieved using backpropagation, which efficiently computes gradients by exploiting the chain rule between layers of the model. The gradient-based optimization method chosen depends largely on the use case.

Only ResNet-18 was used in this report, using pre trained Imagenet1k V1 weights which are commonly used on pytorch and have shown to have good performance.[23] The decision to use pre-trained models was based on the way dataset has been formulated to reduce training times due to the limited compute, discussion of how the data has been formulated can be found in section 5.1. ResNet18 was chosen for its small size and ability to handle vanishing gradients. EfficientNet B0 was tested but unfortunately dramatically increased computation time likely due to the scaling nature of the model.

All models tested in this study were based on a base architecture, illustrated in Figure 7, with variations introduced depending on the specific hypothesis under investigation. This base architecture was adapted from the winning model of the Kaggle competition [4], chosen for its flexibility in handling both tabular and image data, as well as for its relatively shallow depth—an advantage when working with limited data, as deeper networks typically require larger datasets to generalize effectively. Additionally, this base model served as a benchmark for evaluating the performance of modified architectures, enabling a consistent framework for assessing potential improvements. It is important to note that direct implementation of the original competition models would be inappropriate, given the substantial computational resources used in their training, which far exceed those available in the current study.

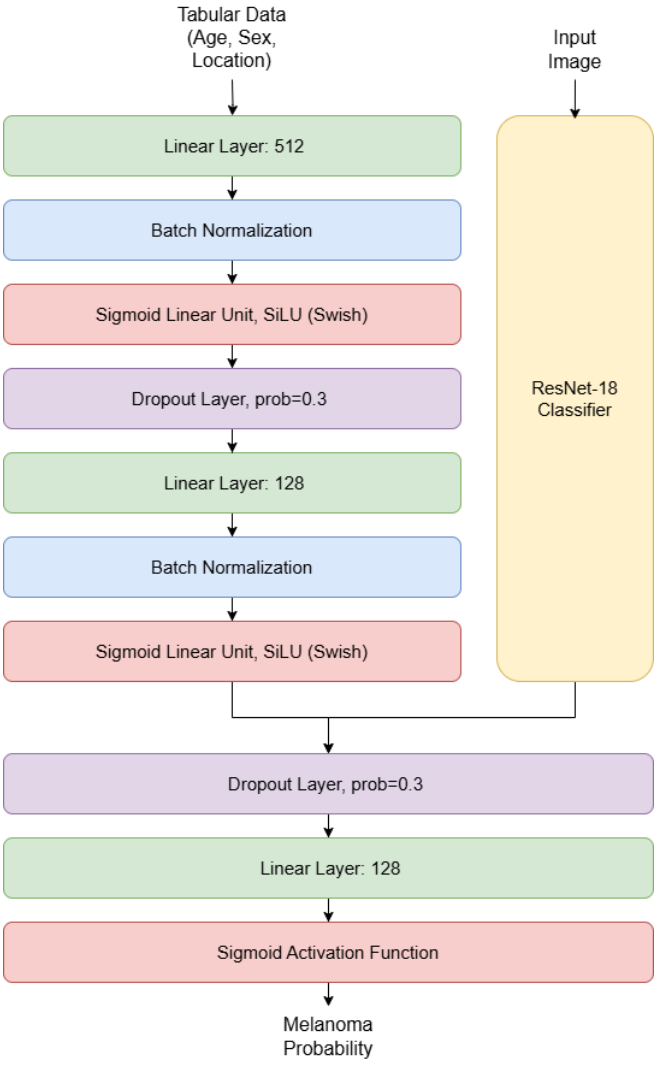


Fig. 7. Base Architecture for Models Tested.

The base architecture differs from the one proposed in [4] in that a binary cross entropy loss function is used due to the singular output as opposed to the cross entropy loss combined with multi output predictor of diagnosis (which was also provided in the dataset but had very little information). This was done to simplify the computation of the ROC-AUC metric which can become somewhat obscured in the multiclass case, as it requires a decision of the one-vs-one or one-vs-all formulation. As a result, the sigmoid activation function is also used, rather than the softmax at the output layer due to the binary nature of the predictor.

The primary difference between the base architecture and that of the additional ones is how the input image is processed and whether the tabular data is used. The first additional model splits the male and female data samples to train separate ResNet image classifiers using the tabular data to assist. The second additional model splits images based on how dark the images, it takes the brightness of the middle images of the melanoma images after they have been sorted by brightness.

As discussed in section 3 the data within the dataset may be better suited for a MoE model that is tailored to sex or image brightness. To my knowledge, none of the solutions on Kaggle explore this aspect, they solely rely on training on an increased amount of data and training on different models of different sizes. Thus, these models intend to explore whether there are performance gains can by designing the models to better suit the dataset.

5. IMPLEMENTATION

It was decided that all of the models would be run locally, partly due to constraints of the assignment but also to explore alternatives to the majority of the entries in the competition which primarily cloud computing for their training over periods of days in some cases.

5.1. Data

To help with storage and computation all images were resized to 256x256 pixels and downsampled to 60% of their original resolution. Larger images were considered but this size was deemed the most suitable for the hardware being used. The data contained duplicate images which were removed along with na values in the dataset. A get_data function was developed to convert the data to the desired format and to new folders.

As highlighted in section 3 it was hypothesized that sampling equally across each of the patients may help to reduce the size of the dataset while also maintaining the diversity of the dataset, which is necessary for generalization. To test this, functionality was built into the code that allowed testing of the models with both a smaller sample of the dataset and the full dataset. This alternative dataset is termed the smaller dataset. This new dataset contains only 3987 images compared with the 33,126 in the original dataset.

The training data was split into validation and training data. In most of the Kaggle competitions they did this by performing k-fold cross-validation. This is computationally very expensive and so the initial training data was split into train and validation data with an 85:15 split. To ensure no data leakage the images from the same patient only appeared in one of the two datasets. For the smaller dataset, all melanoma images were taken from the training dataset before sampling 2 images from each of the patients. The data was then split into train and validation with the same split as before. All of this was performed by sampling the CSV containing the image IDs, patient IDs and tabular data. The "preprocess" function performed all of this with the functionality to choose whether to produce the smaller dataset. Additionally, tabular data was preprocessed using one-hot encoding.

Synthetic data is a common way to increase the amount of available training data. For image data, this is often a simple process by applying common image transformation such as flips, rotations crops and colour adjustments. Alternative more advanced methods include using AI image generators such as diffusion models or Generative Adversarial Networks. Some basic diffusion models were tested for this task but did not produce promising results and due to the uncertainty of efficacy were not explored further. Thus only standard image transformations using the albumentations package including an image flip, shift, scale, rotate and brightness adjustment were considered. However, in testing the dataset with all the images as well as synthetic images were too large to compute in a reasonable amount of time (more then 1 day for a single model). Instead, image transformations were only applied to the small dataset so to increase the number of positive samples available for training without running into complexity issues. Three and five augmentations per image were tested to explore the effects of the transformations.

5.2. Models

Training of the models was carried out using checkpointing to ensure the best model from all epochs is used. The batch size was fixed to 64 to speed up training and the number of epochs was limited to 10 which was more than sufficient for all models, most of the models hit the early stopping criteria of no improvements in validation loss for 3 epochs. The learning rate was initialised at $1e^{-5}$ and scheduled using a cosine annealing schedule which helps the model to converge more quickly. The adam optimizer was used for all models for a similar reason.

The function "train" was developed to apply the early stopping, check pointing and saving of the model as well as validations after each epoch. It includes functions to train the model for each epoch as well as validate after each. The model computes many metrics after each epoch using an ini-

tial set threshold of 0.5, including the F1 score, precision, recall, ROC AUC, and confusion matrix; saving them along with the model weights to a directory.

Hyperparameter tuning of the models such as batch size, initial learning rate, dropout rates and other model specific parameters were not explored significantly due to the training time of the models, over 5 hours for the base model with the complete dataset.

Training and validation of the models was performed separately from testing as Kaggle does not provide the target outputs for the test dataset. Additionally, analysis of the results was performed in a jupyter notebook away from the main training pipeline for flexibility in analysis. Each model type was tested on each of the five datasets, namely: all (all images), small (selection of images), small 3x aug (small with three augmentations) and small 5x aug. The pipeline iteratively passes through each of the preprocessing steps avoiding any redundant transformations by checking for images that already exist.

At the end of the training cycle, the best model from the training batch was evaluated on the training dataset. It should be noted that to ensure a fair comparison of the models, the same validation dataset was used for model comparison. Due to the formulation of the different datasets, this meant that rows needed to be removed from the validation dataset to ensure there was no data leakage. The validation dataset used was that of the "all images" datasets and so by comparing the rows used in training for each of the smaller models, rows could be identified and subsequently removed.

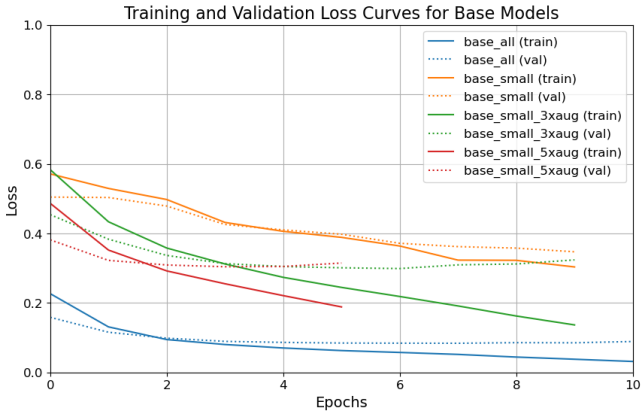


Fig. 8. Validation and Training Loss curves for Base model on four separate datasets.

After training, the loss curves for each of the models was computed, the training curve for the base model can be seen in 8. Additionally, the training curves for the other models can be found in the appendix. This was done to ensure the models were trained correctly. Generally, it can be seen that some of the models underfit to the validation data quite heavily, this is discussed further in section 6.

6. EXPERIMENTAL RESULTS AND ANALYSIS

The initial results showed that the sex-based and brightness-based models were failing to train on the data resulting in ROC-AUC scores that were no better than a random guess. Despite efforts to rectify this, the models continued to underfit the data. The only dataset where the sex-based or brightness-based models did not completely underfit the data was when all the images were used. The ROC for these two models and their corresponding datasets can be found in the appendix in Figures 16 and 17. Likely the reason for this is there is simply not enough data for each of the classes to effectively train both of the image feature extractors.

However, the base model does appear to be training well and produces reasonable ROC-AUC scores across all four of the datasets, as seen in figure 9. Interestingly, the base model trained with the small dataset did not appear to perform significantly worse than that of the model trained with all the data. This is in contrast to the two previous models which both needed more data to learn any sort of predictive features of the data.

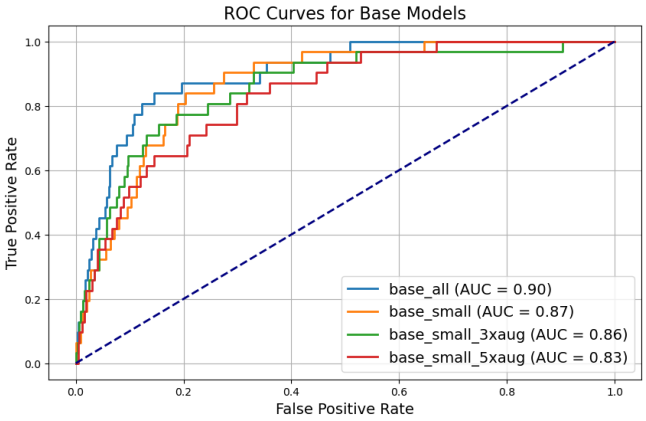


Fig. 9. Base Model ROC for all.

It is likely that the reason for this lies both in the model architecture but also in the information encoded into the data. The base model is essentially twice as small as the sex and brightness-based models. This often results in less data being required to learn predictive features in the data, obviously there is a balance here as too small of a model and it will not be able to learn the more complex features of the data. Likely due to prioritizing the melanoma data within the smaller datasets, the useful information available to the model remains high despite reducing the size of the dataset by nearly an order of magnitude. Also, as samples are still maintained from each of the individual patients, there exists a large diversity in the colour of skin and quality of the image being captured.

It was decided that the base model would only be taken forward at this point. To determine accurate values for the

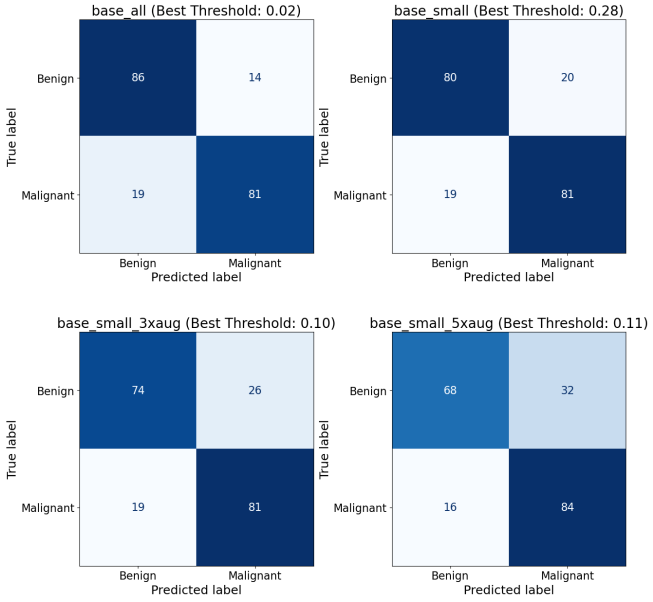


Fig. 10. Confusion Matrix of each of the base models with optimized thresholds applied.

other performance metrics, the threshold was optimized for each of the trained Base models to see how well the classifier is actually performing. A report of the performance of the Base model with the different datasets can be seen in 6 along with the optimal threshold calculated. The optimal threshold was calculated by maximising the F1-score and bounding the true positive rate to between 0.8 and 0.9 to ensure priority for correctly predicting more melanomas.

In it, it can be seen that the model trained with all the data has very low threshold of 0.02. A value this low may cause issues as the precision for predictions needs to be far higher then with a higher tolerance, like that seen in the base model. Additionally, it can be seen that the recall remains constant across all of the tests indicating that they are all equally good at predicting true positives, likely the most important metric for an imbalanced dataset. To better illustrate this, confusion matrices for each of the base models were created, as seen in Figure 10.

Table 1. Table of metrics of the performance of the base model at optimal thresholds as well as the ROC-AUC.

Model	Threshold	F1-Score	ROC-AUC	Recall
Base All	0.02	0.056	0.900	0.806
Base Small	0.28	0.041	0.868	0.806
Base Small 3xaug	0.10	0.033	0.856	0.806
Base Small 5xaug	0.11	0.027	0.834	0.806

The lack of effect of the augmented images is also interesting. It shows that the model is able to learn as many important features with a smaller amount of data, and actually has better performance overall. Likely the reason for this comes from the augmentations causing a slight overfitting to the data, as once the important features are extracted additional data may only result in overfitting unless the model architecture is changed.

Not only did the base model with the small dataset perform remarkably well compared with all the data, it also trained in a far shorter time requiring less than 30 minutes to train compared with over five hours for the base model with all the data. This result reinforces the importance of using data that is high quality and diverse, and not simply numerous.

To better understand where the base model trained with the small dataset was performing incorrectly, six true positive and false negative images have been shown in Figures 11 and 12. Visually, it is very hard to distinguish the difference between these samples, however within the false negatives there are three mole that are generally far smaller then that in the True Positives. This may indicate that the model may not be seeing many examples of small melanomas in the dataset or that the dataset contains far more examples of small benign moles then small malignant ones.

Fig. 11. True Positive Predictions from the Base Small model.

Fig. 12. False Negative Predictions from the Base Small Model.

As with almost all ML problems more data is beneficial, particularly melanoma data points in this case. The model needs to see more examples of all types of melanomas to be able to accurately predict them all. The idea of generalization for CNN models is quite difficult to achieve particularly on very imbalanced datasets like this were there simply isn't enough information in the melanoma samples.

As with all Kaggle competitions, the final model can be ranked against other users on a completely unseen test dataset where for this competition only a ROC-AUC value is given. For this competition the Base model trained on the small dataset achieved a ROC-AUC of 0.845, placing it in the middle of the leaderboard with first place coming in at 0.949.

7. CONCLUSIONS

This study explored alternative approaches to melanoma classification using limited computational resources, challenging the common practice in competitions of relying solely on extensive data and computational power. The primary objective was to investigate whether thoughtful dataset construction and model selection could achieve competitive performance without the extensive resources typically employed in Kaggle competitions.

The most significant finding was that carefully selecting a smaller, more balanced dataset (3,987 images) yielded comparable performance to using the full dataset (33,126 images), achieving a ROC-AUC of 0.868 on the validation set and 0.845 on the competition test set. This represents a substantial efficiency improvement, reducing training time from five hours to just 30 minutes while maintaining competitive performance. The smaller dataset was constructed by strategically sampling across patients to maintain diversity while prioritizing melanoma samples, demonstrating that data quality and diversity can be more important than sheer quantity.

While the specialized mixture of experts models (sex-based and brightness-based) underperformed due to insufficient data for separate feature extractors, this investigation revealed valuable insights about model complexity and data requirements. The analysis of true positives versus false negatives suggested that model performance could be improved by ensuring better representation of small melanomas in the training data.

This research contributes to the field by demonstrating that resource-efficient approaches can achieve respectable results in medical image classification tasks. For practical applications in resource-constrained environments like remote healthcare settings, our findings suggest that optimizing data selection and model complexity may be more beneficial than simply accumulating more data or computational power.

Future work could explore more sophisticated methods for constructing balanced, representative datasets, investigating other potential splits for mixture of experts models, and developing ensembling techniques that maintain computational efficiency while improving overall performance. Additional attention to model interpretability would also be valuable for clinical applications, as trust and transparency remain essential factors in medical AI adoption.

Overall, this project highlights the value of thoughtful dataset construction and model selection in achieving efficient, effective melanoma classification, potentially making such systems more accessible in resource-limited healthcare settings.

8. REFERENCES

[1] Cancer Research UK, "Melanoma skin cancer statistics," 2024, Accessed: 2025-03-24.

[2] Getting It Right First Time (GIRFT), "Dermatology: Girft programme national specialty report," 2021, Accessed: 2025-03-24.

[3] Cleveland Clinic, "Melanoma: Symptoms, stages, diagnosis, treatment, and prevention," 2024, Accessed: 2025-03-24.

[4] Qishen Ha, Bo Liu, and Fuxu Liu, "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge," 2020.

[5] GlobalLogic, "A guide to offline machine learning in healthcare devices," 2024, Accessed: 2025-03-24.

[6] Katharina Sies, Julia K. Winkler, and Christine Fink et al., "Does sex matter? analysis of sex-related differences in the diagnostic performance of a market-approved convolutional neural network for skin cancer detection," *European Journal of Cancer*, vol. 164, pp. 88–94, 2022.

[7] Mingxing Tan and Quoc V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[9] Con Dolianitis, John Kelly, Rory Wolfe, and Pamela Simpson, "Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions," *Archives of dermatology*, vol. 141, pp. 1008–14, 09 2005.

[10] H.A. Haenssle, C. Fink, and R. Schneiderbauer et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018, Immune-related pathologic response criteria.

[11] Gerard Canal, Rita Borgo, Andrew Coles, Archie Drake, Dong Huynh, Perry Keller, Senka Krivić, Paul Luff, Quratul ain Mahesar, Luc Moreau, Simon Parsons, Menisha Patel, and Elizabeth I Sklar, "Building trust in human-machine partnerships," *Computer Law Security Review*, vol. 39, pp. 105489, 2020.

[12] Umang Bhatt, Pradeep Ravikumar, and Jos´e M. F. Moura, "Building human-machine trust via interpretability," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 9919–9920, Jul. 2019.

[13] Mohammadreza Sheykhmousa, Masoud Mahdianpari, Hamid Ghanbari, Fariba Mohammadimanesh, Pedram Ghamisi, and Saeid Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308–6325, 2020.

[14] Melissa D. McCradden and Ian Stedman, "Explaining decisions without explainability? artificial intelligence and medicolegal accountability," *Future Healthcare Journal*, vol. 11, no. 3, pp. 100171, 2024.

[15] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, pp. e190043, 2020, PMID: 32510054.

[16] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[17] Alexander Katzmann, Oliver Taubmann, Stephen Ahmad, Alexander Mühlberg, Michael Sühling, and Horst-Michael Groß, "Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization," *Neurocomputing*, vol. 458, pp. 141–156, 2021.

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[19] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Carola Berking, Sebastian Haferkamp, Axel Hauschild, Michael Weichenthal, Joachim Klode, Dirk Schadendorf, Tim Holland-Letz, Christof von Kalle, Stefan Fröhling, Bastian Schilling, and Jochen S. Utikal, "Deep neural networks are superior to dermatologists in melanoma image classification," *European Journal of Cancer*, vol. 119, pp. 11–17, 2019.

[20] Rajdeep Kaur, Rakesh Kumar, and Meenu Gupta, "Review on transfer learning for convolutional neural network," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 922–926.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] V.H. Phung and E.J. Rhee, "A deep learning approach for classification of cloud image patches on small datasets," *Journal of Information and Communication Convergence Engineering*, vol. 16, pp. 173–178, 01 2018.

[23] PyTorch Foundation, "Torchvision models," 2024, Accessed: 2025-03-26.

9. APPENDIX

Fig. 13. Plot of the distribution of sex data compared with age and mole classification.

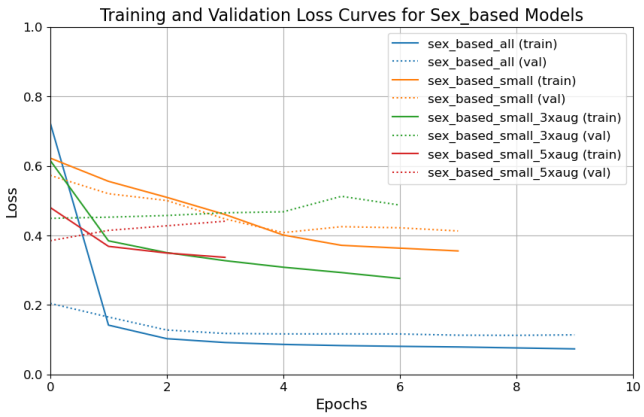


Fig. 14. Validation and Training Loss curves for Sex-based model on four separate datasets.

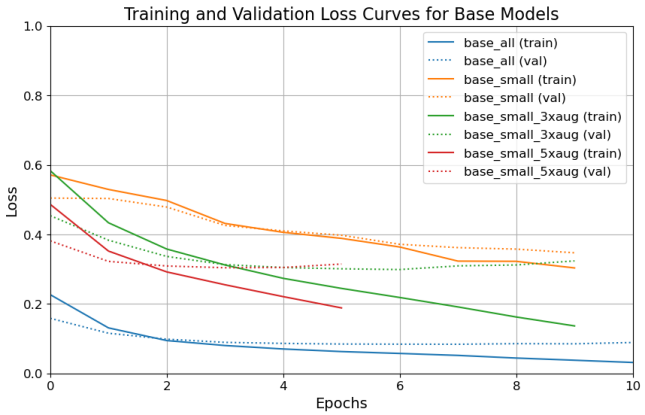


Fig. 15. Validation and Training Loss curves for Brightness-based model on four separate datasets.

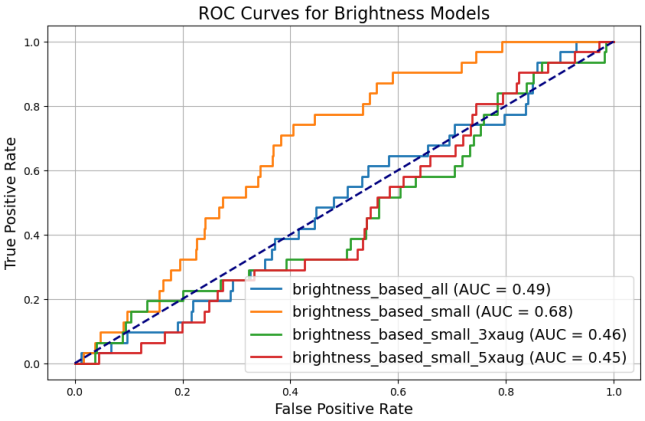


Fig. 16. ROC for the Brightness-based model.

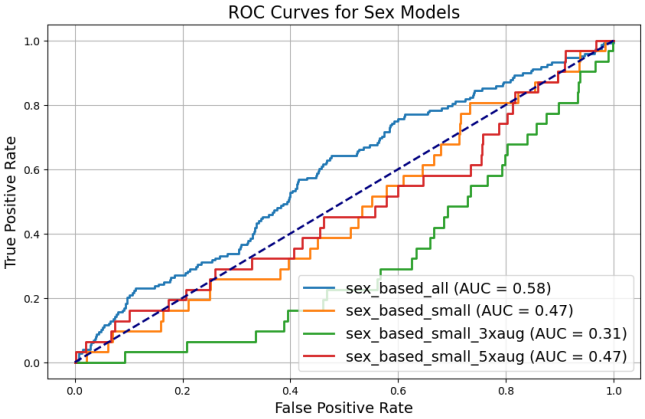


Fig. 17. ROC for the Sex-based models.