Tackling Racial Bias in Machine Learning for Dermatology

D. Aylward
School of Physical Sciences
Dublin City University
darragh.aylward2@mail.dcu.ie

N. E. De Torres
School of Physical Sciences
Dublin City University
niel.detorres2@mail.dcu.ie

K. Ibebugwu
School of Chemical Sciences
Dublin City University
kosisochukwu.ibebugwu2@mail.dcu.ie

R. McCann
School of Computing
Dublin City University
ross.mccann25@mail.dcu.ie

S. McElligott
School of Computing
Dublin City University
sam.mcelligott3@mail.dcu.ie

ABSTRACT

Supervised machine learning methods have demonstrated significant potential in medical applications, particularly in diagnosing dermatological conditions. These models can improve healthcare efficiency by automating diagnostics, but biases in datasets and preprocessing often lead to disparities in performance, particularly for non-white patients. The Stanford Diverse Dermatology Images (DDI) dataset, is a public dataset consisting of 656 biopsy-confirmed [1] dermatological images across diverse skin tones and conditions, providing a benchmark for addressing these disparities. In this project, convolutional neural networks (CNNs) and fine-tuning techniques are deployed to develop a model capable of accurately identifying malignant skin conditions across multiple skin tones. Initial experiments aim to improve evaluation metrics such as ROC-AUC, compared to the results reported in the Stanford study. By enhancing fairness and diagnostic accuracy, this work seeks to contribute to the development of reliable and equitable Artificial Intelligence tools for dermatological care.

Keywords—supervised machine learning, deep learning, Convolution Neural Networks, dermatology, bias, ROC-AUC, sensitivity, Diverse Dermatology Images

I. INTRODUCTION

The advanced progression in machine learning has significantly impacted the medical diagnostic process with its ability to offer automated solutions to problems that typically rely on human expertise in the past. Dermatology can greatly benefit from these advancements due to the dependence on visual diagnoses for skin conditions. Biases in training datasets and limitations in algorithms, result in health disparities often present with patients with darker skin. These disparities highlight the urgent need for models that can fairly and accurately diagnose all skin tones.

This project builds upon the work previously established in the study by Chen et al. titled 'Disparities in Dermatology AI: Assessments Using Diverse Clinical Images' which highlights the limitations of existing dermatology AI models when applied to diverse skin tones [2]. This study introduces the Diverse Dermatology Images (DDI) dataset. This is a public dataset consisting of 656 dermatology images which encompasses multiple Fitzpatrick Skin types (I-VI) [1]. The dataset comprises 656 biopsy-confirmed dermatology images, with labels for condition and malignancy status stored in the metadata. These annotations form a ground truth for evaluating model performance. By including images across multiple Fitzpatrick skin types (I–VI), the dataset provides a reliable benchmark for identifying and addressing racial disparities in dermatological Artificial Intelligence (AI). While the Stanford study demonstrated performance improvements through fine-tuning, significant gaps remained in model accuracy and fairness, particularly for underrepresented skin tones.

The goal of this project is to enhance the performance of machine learning models in identifying malignant skin conditions across diverse skin tones. Convolutional Neural Networks (CNNs), pre-trained on large datasets and fine-tuned using the DDI dataset, are employed to achieve this objective. By optimising evaluation metrics such as ROC-AUC, F1-Score and sensitivity, this work aims to exceed the results reported in the Stanford study, contributing to the development of reliable and equitable AI tools for dermatological care.

A. Dataset Summary

The DDI dataset consists of 656 biopsy-confirmed images labeled with corresponding skin conditions and malignancy status [1]. Images are categorised by Fitzpatrick skin types, ensuring representation across diverse skin tones. Each image includes metadata specifying key attributes like diagnosis, malignancy, and skin tone, facilitating stratified analysis. To address dataset limitations, data augmentation techniques and class balancing was

applied. Despite its strengths, the small dataset size and class imbalance pose challenges for deep learning applications.

II. Algorithms

This project utilises transfer learning and uses both TensorFlow/Keras and PyTorch frameworks to classify dermatological images. The methodology is centred around optimising the performance of deep learning models on the Diverse DDI dataset. This section details the preprocessing pipeline, model architecture, and training strategy used throughout.

A. Preprocessing

1) Class Balance

Oversampling is introduced to address the class imbalance present in the dataset, increasing the proportion of malignant samples to match that of benign samples. This ensures the model is trained on a balanced dataset, providing accurate and consistent learning [3].

2) Data Augmentation

Augmentation enhances the model's ability to generalise across the dataset. Techniques involve; random flips, rotations (±20°), zooming (±20%), and contrast adjustments (±10%). These techniques are only applied to the training set as the testing data needs to remain unbiased for fair evaluation.

3) Stratified Splitting

An 80/20 train-test split is performed with stratification based on skin tone and malignancy labels. from the 80% training set split, 10% of this was used for validation. This ensures proportional representation across subsets, improving fairness and consistency.

4) Data Preparation

Images are resized to 300x300 pixels with padding to maintain aspect ratio and normalized to a [0,1] range. Datasets are batched and pre-processed for efficient training using *tf.data.Dataset* in TensorFlow or *DataLoader* in PyTorch.

B. Model Architecture

ResNet152 is a deep CNN used as the base model for transfer learning in this project. ResNet152 has 152 layers which are capable of extracting key features and patterns making it optimal for parsing through the images in the DDI dataset [4].

1) Pre-trained Layers

ResNet152 is initialized with weights trained on ImageNet, leveraging its learned features for medical image classification. We reused its convolutional layer and its Global Average Pooling (GAP) layer. The purpose of the GAP layer is to reduce the dimensionality of the feature maps by averaging each map into a single value, to be passed as a vector to the fully connected layer [5].

2) Fine Tuning

The earlier layers of ResNet152 are frozen to retain their pre-trained knowledge, while the fully connected layer is modified and retrained for the binary classification task (malignant vs. benign). We changed the fully connected layer to a sequential block which maps the output of the GAP layer to 128 outputs, applies a Rectified Linear Unit (ReLU) activation function and Dropout layer, maps the 128 outputs to a single value then finally applies a Sigmoid function to obtain a value between 0 and 1 (predicted probability).

3) Loss Function & Optimisation Strategy

We used a Binary Cross-Entropy loss function as our criterion for guiding the training process because it is suitable for binary classification tasks. We used an Adam optimiser due to its efficiency and its suitability for binary classification tasks. We trained with a learning rate of 0.0008 after determining it to be suitable at preventing overfitting.

III. Challenges and solutions

A. Pre trained model selection

Our first idea was to use a model trained on an existing dermatology image dataset, e.g. HAM10000 [6]. This proved troublesome however, as the model we found had specific layer input and output expectations, which we could not figure out. We settled on using ResNet152, which offered the best performance and was easy to implement.

B. Library selection

We originally wrote our training algorithm using Tensorflow, but its implementation of the ResNet family of models resulted in poor performance with the DDI dataset. We also found the Tensorflow documentation hard to follow, so we rewrote our code using PyTorch and this solved both problems.

C. Dataset imbalance

The DDI dataset is incredibly imbalanced, with it containing 485 benign cases and only 171 malignant ones. Our solution was to implement oversampling on the minority class (malignant = True), which combined with our data augmentation pipeline allowed us to balance the data while avoiding overfitting. We also implemented stratification on the *skin_tone* and *malignant* labels

in order to maintain a balanced proportion of samples in the training, validation and testing splits.

D. *Overfitting*

Our first few implementations were very prone to overfitting (due to the dataset being relatively small), which we monitored by keeping track of validation loss and training loss. To fix this, we adjusted our learning rate in order to maintain a steady decline in validation loss. We also implemented a dropout of 0.5 which ensured biases learned by the model would be removed on each training iteration.

IV. Evaluation Metrics

A. *Accuracy*

Accuracy is by far the most popular and relevant metric for measuring the performance of a machine learning model. A simple measurement, classification accuracy is defined as the number of correct predictions that a model makes divided by the number of total predictions that a model makes. In section II A-2, we discussed using oversampling and data augmentation to balance the dataset. The classification accuracy metric is best used to represent model accuracy when the dataset is balanced [7], thus justifying its use as a metric in our evaluation.

B. *ROC-AUC*

The area under the receiver operating characteristic curve (ROC-AUC) is another measurement used for classification machine learning models to evaluate performance. The ROC is a probability curve, and the AUC represents the degree of separability. What the ROC-AUC result says, is how good a machine learning model should be at distinguishing differences in classes [8]. The closer to 1 that the ROC-AUC is, the better that model should be at classification. If the ROC-AUC is above 0.5, it represents that the model is at least better than guessing. It should be noted that the ROC-AUC measure tends to be a better evaluation metric on imbalanced datasets, but as essentially every reputable paper on machine learning that we read includes it in their evaluation, we decided to do the same, as it can still provide useful information about the model, even on a balanced dataset.

C. *F1 Score*

The F1 score is described as the harmonic mean between precision and recall. Precision shows how many positive classifications the model made that are actually positive, described as the number of correctly classified positives, divided by the number of everything classified as positive. Recall is a similar evaluation metric that shows how a model correctly identifies true positives. Recall is described as the amount of correctly classified actual positives, divided by the number of all positives. The F1 score tries to find the balance between precision and recall [7]. We decided to use the F1 score as an evaluation metric, as it is usually a good indicator of how robust a model is.

D. *Confusion Matrix*

A confusion matrix gives a very easy to understand graphical representation of the performance of a model. Results of the performance are shown in a 2x2 grid, describing the number of false positives, false negatives, true positives and true negatives. We decided to use a confusion matrix as an evaluation metric due to this simplicity, giving us a quick and easy way to view the overall results of our model.

V. Results & Discussion

A. *Model Learning Progression*

After fine tuning, the final hyperparameters used for the model were a learning rate of 0.0008 and 15 epochs. Both training and validation loss were recorded for each epoch. The plot of loss against each epoch is shown in Fig. 1 and shows a clear downward trend which indicates effective learning and good generalisation. This proves that the model successfully avoided overfitting or underfitting. The plateau for validation loss seen in later epochs indicates that the model has learned all it can from the data.

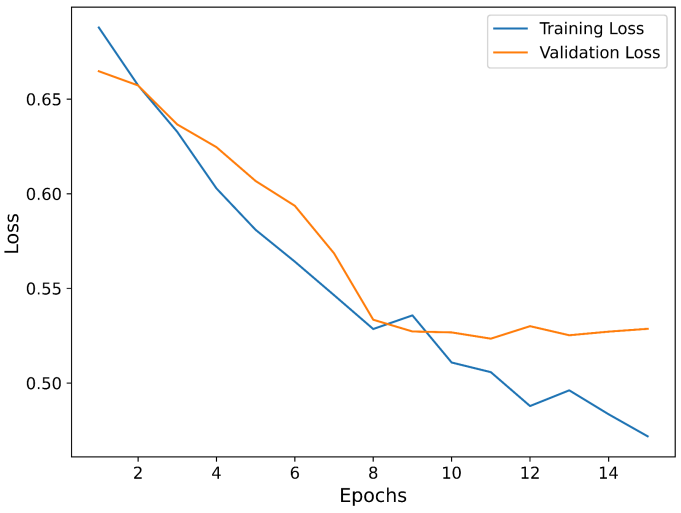

Fig. 1 The loss of the model plotted against epoch.

B. *Confusion Matrix*

The confusion matrix shown below in Fig. 2 clearly demonstrates the model's performance for the two classes: benign and malignant. The model shows a strong ability to identify benign cases, with an accuracy of 86% while its performance in identifying malignant cases was slightly lower at 74%. The overall accuracy was 80% which indicates reasonably high performance but also room for improvement, particularly in malignant case detection. The discrepancy between benign and malignant cases is probably due to the fact that there's more complexity and variability involved in detecting malignant lesions.

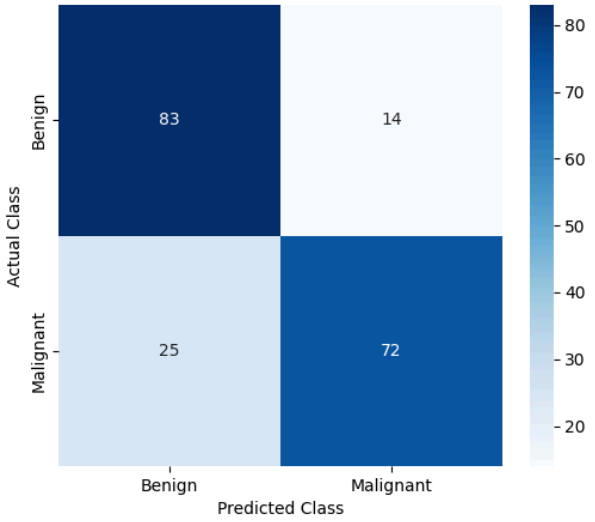

Fig. 2 The Confusion Matrix

C. *Evaluation Metrics*

To quantify performance, a range of metrics were calculated and shown below in Table 1. The ROC-AUC score of 0.86 demonstrates that the model is very effective at distinguishing between the two classes. The balanced F1 scores for benign (0.81) and malignant (0.79) cases highlight the model's robustness. However, the disparity between recall for benign (0.86) and malignant (0.74) cases suggests that the model is slightly biased toward identifying benign cases. This is a potential limitation that could be addressed through further model refinement.

Table 1. Performance metrics for the model.

Metric	Benign	Malignant	Overall
Accuracy	0.86	0.74	0.80
ROC-AUC	-	-	0.86
Precision	0.77	0.84	0.80
Recall	0.86	0.74	0.80
F1 Score	0.81	0.79	0.80

VI. Conclusion

It is evident from Table 1 in the previous section that the transfer learning approach to malignancy detection is certainly feasible, and that with further research can become a life saving application of the technique. The discrepancy between benign and malignant prediction accuracy indicates that there is more nuance in detecting malignant skin lesions than expected, but we believe that this is something that can be improved upon with larger, higher quality datasets. This model's ability to perform relatively well on diverse skin types also suggests that with proper techniques, racial bias can be minimised in the field even further.

VII. References

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci Data*, vol. 5, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.

[2] R. Daneshjou *et al.*, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Sci. Adv.*, vol. 8, no. 32, p. eabq6147, Aug. 2022, doi: 10.1126/sciadv.abq6147.

[3] J. Gong, "A Novel Oversampling Technique for Imbalanced Learning Based on SMOTE and Genetic Algorithm," in *Neural Information Processing*, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds., Cham: Springer International Publishing, 2021, pp. 201–212. doi: 10.1007/978-3-030-92238-2_17.

[4] H. R. Fahim, *HR-Fahim/ResNet-152-Model-Testing-with-PyTorch*. (Mar. 13, 2024). Python. Accessed: Jan. 22, 2025. [Online]. Available: https://github.com/HR-Fahim/ResNet-152-Model-Testing-with-PyTorch.

[5] [1] Y. Dogan, "A New Global Pooling Method for Deep Neural Networks: Global Average of Top-K Max-Pooling," *TS*, vol. 40, no. 2, pp. 577–587, Apr. 2023, doi: 10.18280/ts.400216.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, Aug. 2018, doi: https://doi.org/10.1038/sdata.2018.161.

[7] A. Mishra, "Metrics to Evaluate Your Machine Learning Algorithm", Medium, https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (accessed 22/01/2025).

[8] S. Narkhede, "Understanding AUC - ROC Curve", Medium, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (accessed 22/01/2025).