Utilizing Data Augmentation to Assist Melanoma Detection Models Across Diverse Racial Groups

Nam Tran
East Hamilton High School

Current machine learning (ML) models developed to tackle the issue of melanoma detection struggle to detect melanoma in diverse skin groups, namely individuals from African and Hispanic backgrounds. This problem primarily stems from a lack of diverse data available to train ML models. As a solution, this paper introduces the use of data augmentation to correct both a lack of data diversity and class imbalances. For this study, data augmentation was applied to better enhance the performance of two ML models for diverse melanoma detection: ResNet50 and Random Forest. Using the state-of-the-art HAM10000 and Diverse Dermatology Images (DDI) datasets, the researcher examined how well each model did in distinguishing melanoma from both the lighter skin in the HAM10000 dataset and the darker skin in the DDI dataset before and after data augmentation was applied. When assessed, both models across both datasets saw improvements across all metrics, especially recall. Overall, however, ResNet50 (accuracy = 0.55) and Random Forest (accuracy = 0.44) obtained significantly worse results when shown the darker-skinned images. Thus, while the results reaffirm the practice of data augmentation to correct poor class balances and to increase the diversity of training data, its use was not able to reduce the disparity gap between Caucasian and darker racial groups in terms of melanoma detection.

Key Words: melanoma detection, data augmentation, machine learning, racial disparities, ResNet50, Random Forest, skin cancer classification

Introduction

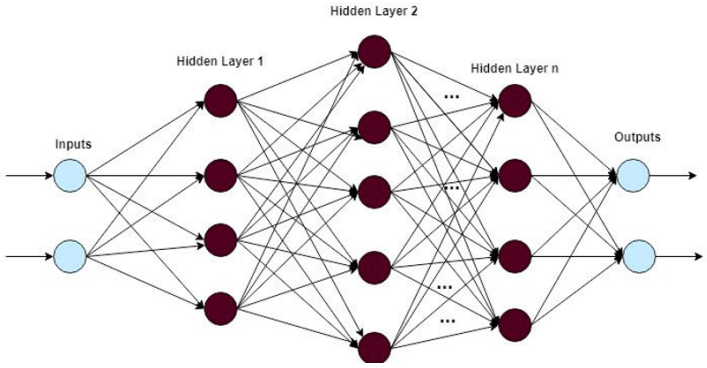
Background

Melanoma is the deadliest skin cancer despite its low risk and rarity. Despite only accounting for 1% of all cases of skin cancer (American Cancer Society, 2024), melanoma is still the leading cause of death over all other skin diseases (Babar et al., 2021). Moreover, melanoma has an abnormally rapid rate of cell multiplication (Thurnhofer-Hemsi & Domínguez, 2020), becoming deadlier over time. In fact, patients in the later stages of melanoma (Stage IV) only have a 10-15% chance of

surviving for ten years, compared to the 94-98% survival chance they have during earlier stages (Stage I) (Bhatt et al., 2022). As a response, many non-invasive methods of diagnosing melanoma early, such as dermoscopy, have been utilized to tackle the task of early detection of melanoma. However, cost constraints and the generally limited pool of expert dermatologists have made early detection of melanoma and the use of non-invasive methods for a majority of clinics and hospitals implausible (Ravi, 2022). Overall, the current state of diagnosing melanoma has resulted in varying diagnoses that differ depending on the dermatologist and their predetermined biases.

Physically, melanoma can be reflected through spots of skin that have darkened significantly in color. However, this appearance change can also be seen in noncancerous forms of skin lesions. A nevus, more commonly referred to as a mole, is a benign skin lesion deriving from melanocytes that also causes the skin to become darker on the surface (American Cancer Society, 2024). Such physical similarity between a healthy mole and a cancerous skin lesion thus calls for more in-depth methods of diagnosing melanoma accurately.

One solution and response to the adversities mentioned previously is the use of artificial intelligence (AI) to automate melanoma detection by taking advantage of their superior feature analysis. In the broad scope of AI, machine learning (ML) is the use of algorithms to create automated machines designed to perform tasks that typically require the intelligence of a human (Alowais et al., 2023). One specific group of ML algorithms that have become widely used and studied is artificial neural networks. Neural networks are a type of machine learning algorithm that model the human brain and the way it learns. The diagram below in Figure 1 displays how the neurons of the nervous system are replicated through nodes. Additionally, similar to how different neurons have different thresholds of activation, the nodes of neural networks also have weights that change as the model trains, a process called backpropagation (Goz et al., 2019). Though this is the general structure of a neural network, additional layers with varying functions can be added or stripped depending on the task at hand. Due to their in-depth learning abilities and deep pattern recognition, artificial neural networks have been classified as a subset of machine learning called deep learning (DL).



Note. This diagram was taken from a research paper written by Choudhary & Kesswani in 2020.

FIGURE 1. Block diagram of a neural network.

While several studies have already shown the use of ML and DL to be successful in detecting cases of melanoma at a high accuracy rate, very few of these models have been developed and trained using images of melanoma in darker racial groups due to a lack of existing diverse datasets (Patel et al., 2023). Hence, the major implication here is that there is a lack of reliable and high-performing models for helping with the diagnosis of melanoma and individuals in all racial groups, including those specifically belonging to African and Hispanic heritage (Daneshjou et al., 2022). If left unaddressed, there are several real-world implications and consequences that occur. Substandard decisions and diagnostics can underestimate or completely disregard the needs of minority groups (Mittermaier et al., 2023), perpetuating the disparities that already exist between various socioeconomic groups clinically (Cross et al., 2024). Resources can be unfairly distributed heavily against groups due to biased data making it so that the already exceeding healthcare costs become seemingly insurmountable to some (Obermeyer et al., 2019). Acknowledging these potential ramifications, the overarching goal and purpose of this study is to improve the current status of melanoma detection models for all racial groups. This is important to help standardize the use of AI and ML technology for diagnosing melanoma in a professional healthcare setting.

Literature Review

Diverse Datasets

As previously stated, current work using ML and DL for assisting in the diagnosis of melanoma lacks the ability to generalize and account for all racial groups (Patel et al., 2023). This stems from these models' biases towards diagnosing melanoma in lighter skin, a bias that originates from the lack of diverse datasets available. Statistics from the American Cancer Society show that White individuals are five to ten times more at risk of contracting melanoma compared to Black and Hispanic individuals (American Cancer Society, 2024). As a result, the amount of detailed and balanced datasets with varying skin complexions for learning models to

train on is relatively scarce, stunting the quantity and quality of machine learning models that can accurately classify melanoma among a diverse population. However, this is not to say that there have not been any advancements made. One substantial step forward is the published Diverse Dermatology Images dataset curated by Daneshjou et al., in 2022, containing an array of a few hundred various skin disease cases, including melanoma, from individuals of darker racial groups. Still, this is incomparable to the many dermatology datasets containing up to tens of thousands of images of lighter-skinned individuals.

Data Augmentation

Dataset-related issues, such as small size, heavy class imbalances, and poor diversity can lead to models inheriting unwanted biases in their classifications (Gong et al., 2023). One solution to these problems is data augmentation. Data augmentation is the process of generating new, unique images by altering the existing images through transformations (Hassan et al., 2022). Such transformations include cropping, zooming, rotating, flipping images, and more (Tong et al., 2020). This method works because computers see an image and its augmentation as two distinct images. Data augmentation is commonly used to address the problems previously mentioned by increasing the number of images available. Specifically, this means that learning models have more data to train with (Morid et al., 2021). Additionally, class-specific data augmentation can prevent models from developing a bias towards classifying a certain class (usually the majority class) by improving the ratio and balance between classes in a dataset (Mumuni & Mumuni, 2022). Lastly, data augmentation also increases the diversity of the data the models are trained on (Aggarwal et al., 2022), which generally means how many unique images there are in a dataset. Overall, data augmentation is a machine learning technique that provides multiple data-specific benefits that helps improve the training of learning models.

Current Findings

Three primary works influenced this research study. Khadija Nawaz et al. leveraged the state-of-the-art FCDS-CNN neural network, a model that applies data augmentation and class weighting techniques to overcome the issues of imbalanced datasets. They evaluated this to other neural networks without the use of data augmentation. The FCDS-CNN model showed stronger classification performance over other comparison models, indicating it to learn more defining features (Nawaz et al., 2025).

Abdelkader Alrabai et al. assessed the use of residual learning in the ResNet50 model, along with three other pre-trained models in the task of skin cancer classification. The ResNet50 model outperformed all the other models across all evaluation metrics used, showcasing its exceptional reliability at distinguishing features between malignant and benign

images. Unlike Nawaz, Alrabai did not address any class imbalances in their datasets with data augmentation (Alrabai et al., 2025).

Sajib A. Dip et al. investigated the use of skin detection models and neural networks previously trained on datasets of lighter skin on the Diverse Dermatology Images dataset, one of the first datasets specifically including images of individuals of darker racial groups. Generally, all the models proposed in this study presented a weaker ability to accurately classify different types of skin lesions in the DDI dataset when compared to the models mentioned in the previous studies. This underscores the performance disparity among melanoma detection models that has been caused by a current lack of highly diverse skin datasets (Dip et al., 2022).

Summary

Holistically, the three aforementioned bodies of work display a brief summary of the current advancements made surrounding the automation of melanoma detection using ML models and deep learning algorithms. However, what each of these studies failed to consider is how the use of data augmentation and neural networks can help correct data-related issues to help improve melanoma detection in more diverse racial groups. While Dip et al., did evaluate existing models on the diverse images of the DDI dataset, he did not specifically examine how the models and techniques by Nawaz and Alrabai could help overcome both the lack of images available nor the distinct differences between skin lesions in light and dark skin groups. Thus, this implores the ongoing question of: How effective is data augmentation in improving the performance of machine learning models in melanoma detection in diverse skin groups? The study aims to combine the techniques and models used by Nawaz and Alrabai to assist an assortment of both DL neural networks and traditional ML algorithms in correctly classifying melanoma skin in darker skin types. This study will be approached with a comparative computational methodology using existing datasets and code written by the researcher.

Methodology

Design

The purpose of this research study was to assess the effectiveness of data augmentation in improving the accuracy of machine learning models designed for melanoma detection in a diverse range of skin groups. Thus, a quantitative comparative computational ex post facto research design was employed to adequately evaluate the before and after.

For image-based ML studies, factors such as cost and accessibility to high-quality medical imaging technology can prevent the ability for personal data collection. The most common response to this problem is deploying an ex post facto research design. Ex post facto is a research design in which the “investigation starts after the fact has occurred” (Silva, 2010); in other words, the data used is pre-existing. Ex post facto was

utilized to take advantage of two already-present dermatology datasets: the HAM10000 dataset and the Diverse Dermatology Images dataset. These datasets will be discussed in greater detail in the Datasets portion.

To advance machine learning's role in the medical field without jeopardizing the health and safety of patients, research studies must be conducted in a way that does not rely on access to medical patients in real time while still being able to produce results predictive of those in real-world scenarios and usage (Adlung et al., 2021). Computational experiments help to fulfill both requirements. A computational experiment is a theoretical research design particularly conducted to "simulate" real-world outcomes in a manner that is "easy to manipulate and repeat" (Wei & Cheng, 2012). Combining ex post facto and computational experiments enables researchers and data scientists to trial varying ML models for medical image-based classification problems, such as the ones in this study.

A quantitative method was used to analyze the image-based data. For computational tasks like data augmentation and feature extraction to be performed on images, the data must be converted into a form that computers can work with: numbers. In a process called image digitization, data in the pictorial form can be converted to numerical data. This is due to the way in which color and pixels must be represented by a computer, in which a single pixel holds three values representing the amount of red, blue, and green there is in the pixel (Wah & Htwe, 2020). In numerical form, computers and algorithms are then able to analyze an image and its patterns. Thus, a quantitative research method was used to both analyze the images and evaluate the performance of the proposed models and algorithms through quantitative metrics.

Datasets

Two image datasets, HAM10000 (Tschandl et al., 2018) and the Diverse Dermatology Images (DDI) (Daneshjou et al., 2022), were utilized in this study. Neither dataset contained any personally identifiable information from any of the images, and both datasets were curated after approval from their respective ethics and review board (University of Queensland; Stanford Institutional Review Board). The HAM10000 dataset is composed of 10,015 professionally taken dermatoscopic images, covering a large assortment of cases of pigmented skin lesions. The individuals represented in this dataset primarily have a lighter skin tone. Comparatively, the DDI dataset is significantly smaller in number, having only 656 images taken photographically. The difference in the methods by which the images of both datasets were taken may interfere with the performance of the models tested. With the images of the DDI dataset being taken by hand, many of the images contain poor quality and objects in the frame beyond just the area of interest on the skin. Conversely, the dermatoscopic images in the HAM10000 dataset were taken under ideal lighting with only the area of interest included. In other words, while the

models will be trained solely using the images from the HAM10000 dataset and augmentations of them, they will also be tasked with classifying images from the DDI dataset that have noise that can interfere with the models' accuracy due to the models picking up on details that have no connection to whether or not a skin lesion is benign or malignant. Regardless, the value of the DDI dataset comes from its composition of darker, diverse skin tones. The inclusion of the DDI dataset was pivotal in the researcher's goal of assessing the effectiveness of data augmentation in improving ML model performance in detecting melanoma among a diverse set of images including darker-skinned individuals.

The researcher filtered through both datasets to isolate cases of melanoma and benign nevi healthy skin lesions. In the HAM10000, only images belonging to the “mel” class (melanoma), “bkl” (benign keratosis) and “nev” class (benign melanocytic nevi) were extracted, accounting for 8917 images. The researcher then separated the filtered images into two classes, with both “bkl” and “nev” images chosen to represent the benign cases and “mel” images representing malignant cases. Table 1 depicts three images that represent the three classes extracted from the dataset.

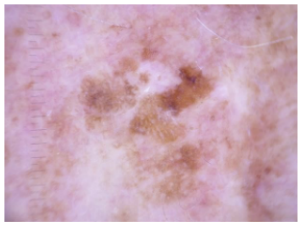
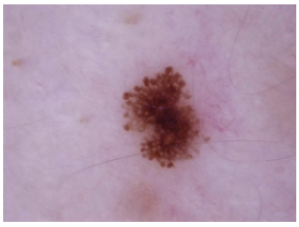
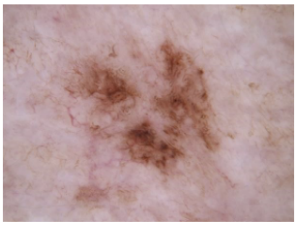
Benign Keratosis	*Melanocytic Nevi*	*Melanoma*
		

TABLE 1. Images from HAM10000 representing the three types of images extracted.

After extracting the ideal images needed for the study, the researcher split the images into a training set and a testing set, using an 80:20 training/testing split. As seen in Table 1, there is a substantial class imbalance, with the benign class overpowering the malignant class. The researcher later addresses this issue using data augmentation.

Subsets	**Benign**	**Malignant**
Training	6243	890
Testing	1561	223

TABLE 2. HAM10000 dataset statistics before data augmentation

In the DDI dataset, images belonging to “melanoma”, “melanoma-in-situ”, “melanoma-acral-lentiginous”, and “nodular-melanoma” (different types of melanoma) make up the melanoma class for this dataset. Images belonging to the “melanocytic-nevi” compose the benign images class for this dataset. These specific images were extracted in accordance with the researcher's specific goal of examining how well each model could differentiate melanoma from images of visually similar images of benign skin lesions. Due to its small size as shown in Table 2, the DDI dataset was only used as a testing set.

Benign	Malignant
119	21

TABLE 3. DDI dataset statistics

Data Augmentation

After extracting and portioning the datasets into training and testing sets, the researcher performed data augmentation on the minority melanoma class to address the class imbalances apparent in the HAM10000 dataset. As previously mentioned, data augmentation helps to counteract shortcomings in datasets by generating new and unique images by transforming the already-available images (Hassan et al., 2022). The researcher performed data augmentation on the mere 890 malignant images in the training set, performing a series of horizontal and vertical flips, rotations, and zooms, improving the diversity of the data and reducing the risk of overfitting, a problem that occurs when a model has overtrained on the training set, causing it to perform poorly when tested against unseen data (Berrar, 2019). All these changes can be seen in Table 4.

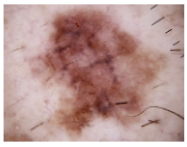
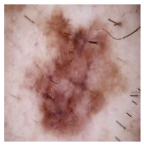
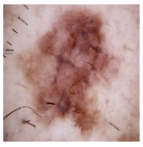
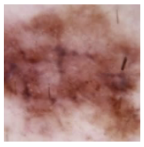
OriginalImage	HorizontalFlip	VerticalFlip	Rotationand Zoom
			

TABLE 4. Melanoma image ISIC_0024482 after various data augmentation techniques

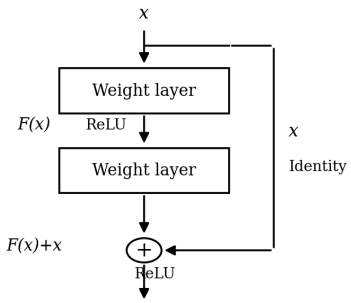
Not only did this drastically improve the diversity and quality of the training data used, but this process helped increase the number of malignant images used four-fold as shown in Table 5, bringing the class ratio closer to 2:1. While there is not a standard or accepted threshold by which a dataset becomes balanced, a majority of public research surrounding imbalanced datasets generally focus on imbalance rates that are 3:1 or greater (Wainer, 2024). This was a significant and sufficient improvement compared to before augmentation. Once again, because the DDI dataset did not have a sufficient amount of images necessary to create a training split, no data augmentation was applied to the darker-skinned images. Moreover, the researcher reasoned that even with the HAM10000 having only images of lighter-skin individuals, the improved diversity and balancing of the dataset would still be able to impose an improvement in melanoma classification in both lighter skin and darker skin.

Subsets	Benign	Malignant
Training	6243	3560
Testing	1561	223

TABLE 5. HAM10000 dataset statistics after data augmentation

Models

The researcher chose two models to evaluate and compare. The first model is ResNet (He et al., 2015). This model was designed to address the problem of the vanishing gradient in deep neural networks. The vanishing gradient is a problem that occurs during backpropagation in which the weights of the nodes at the top of the neural network are unaffected by any learning occurring farther down the neural network. This problem worsens the deeper a neural network is, or the more layers a neural network has (Hochreiter, 1998). The problem originates from the fact that standard neural networks are not able to sufficiently generalize the features of images with deeper network depth. The ResNet architecture combats this issue by using skip connections. Skip connections conceptually work off the idea that the input image of a layer is added onto the resulting transformation of that same input as a way of preserving the features and information of an image for the deeper layers to learn from (He et al., 2015). In other words, the features of the original image are repetitively re-added to the running image undergoing transformations to preserve the original features of the image being analyzed. This reduces the problem of vanishing gradient and thus enables the design of deeper neural networks with more complex layers, allowing a better representation of image features and better classification accuracy.

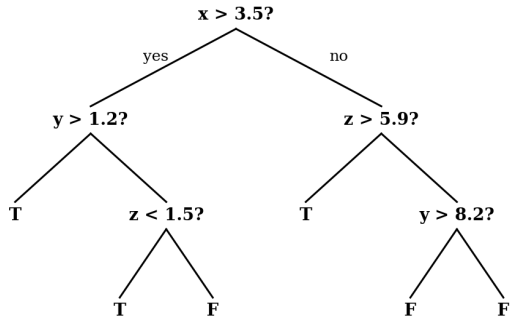


Note. This diagram was taken from a research paper written by Chen et al. in 2020.

FIGURE 2. Block diagram of the skip connection for neural networks

With the inclusion of the residual block, the problem of information being lost in deeper neural networks is significantly negated. Therefore, this model can better generalize more features of an image without losing information (Zhang et al., 2023).

Finally, the Random Forest algorithm is an ensemble classification model invented by Leo Breiman in 2001 (Breiman, 2001). The Random forest is an algorithm that uses groups of decision trees to help produce a single output classification. Similar to flowcharts, decision trees have a hierarchical decision-making structure using a series of descending nodes and branches.



Note. This diagram was taken from a research paper written by Chen et al. in 2020

FIGURE 3. Architecture of Decision Trees

Each different decision tree in the algorithm is responsible for learning a specific subset and a subgroup of features in the data. In other words, each decision tree is only responsible for accounting for certain image features, such as color or shape, and only a certain amount of data when performing classification individually. While this means that a decision tree can help analyze how indicative certain attributes are of certain diseases, the classification accuracy of a single decision tree is not robust due to its easily being overfitted. Thus, a multitude of decision trees are utilized, taking advantage of ensemble learning, in which multiple models are used to enhance predictive performance (Khan et al., 2024).

Training and Hyperparameters

Once coded and compiled, the models were trained on the HAM10000 dataset. Both models were first trained on the version of HAM10000 that did not include the additional malignant images produced from data augmentation. Separate identical models were trained and evaluated on the HAM10000 and DDI datasets after data augmentation.

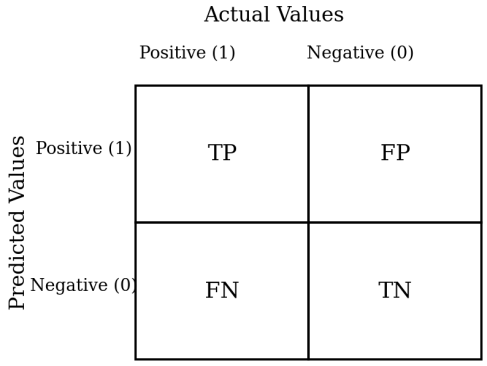
Different hyperparameters and training conditions were used for each model. All the images used to train the models were resized to fit a 224 x 224 size, as per the standard input dimension of ResNet50. For training time, ResNet50 was set to train for 50 epochs (times the model reiterated through the training data (Ajayi & Ashi, 2023)) to ensure that the weights of the model were properly calibrated. Similarly, Random Forest was trained through the training set up to 100 times, each time adding an additional decision tree, though fewer iterations may have happened due to the model peaking in performance early on. To avoid the models from overfitting, cross validation was implemented. Cross validation is a data resampling technique that uses validation sets during training to measure how well the trained models do when shown new data (Berrar, 2019).

Additionally, early stopping was used to save the weights of the models at their peak performance. This was done by tracking when validation loss, a measure of how well a model generalizes to unseen data (Li et al., 2024), fails to improve in consecutive 3 training iterations and stopping training. The model would then save the weights of the model at the state in which validation loss was at its minimum, which would theoretically be when the model reached its peak performance during training. To minimize the computational power needed during training, a batch size of 32 images was used to train the ResNet50 model, while the Random Forest model did not use a batch size for training.

Results

Confusion Matrix

After training the proposed models on both datasets, the researcher evaluated the models after their training with the testing sets made from each dataset. ResNet50 and Random Forest were evaluated using the following statistical metrics using values extracted from the confusion matrix



Note. This diagram was taken from a research paper written by Chen et al. in 2020.

FIGURE 4. Confusion Matrix

The confusion matrix is a table of values utilized for evaluating machine learning models and their performance. In this study, positive values will represent images that the models classified as malignant, while negative values will represent images that the models classified as benign. Moreover, true and false indicate whether that classification was correct (Singh et al., 2021). For example, a False Negative classification represents an instance in which a model incorrectly classified an image as benign.

Metrics

In this study, the researcher attempted to answer the research question: How effective is data augmentation in improving the performance of machine learning models in melanoma detection in diverse skin groups? To decisively answer this question metrics were used to evaluate the performance of the models in this study: accuracy, precision, recall, and F1-Score.

Accuracy: The accuracy metric is calculated by dividing the total number of correctly classified images by the total number of images. Accuracy places equal importance on both classes of the image datasets (Ravi, 2022).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision: Also referred to as the positive predictive value, precision is the ratio of correctly predicted cases of melanoma (TP) to the total number of images a model classifies as melanoma (TP + FP), whether they were correct or not. In other words, precision measures how well the models did at correctly identifying cases of melanoma. With this metric, the researcher can examine how trustworthy a melanoma classification made by the model is (Fränti & Mariescu-Istodor, 2023).

$$Precision = TP / (TP + FP)$$

Recall: The calculated ratio of the number of melanoma images correctly classified by the machine learning model to the number of all melanoma images present in the dataset. This metric measures how many cases of melanoma the models were able to correctly identify out of the total number of melanoma cases. So, while precision measures the trustworthiness of a model, recall measures the reliability of a model to catch all positive/melanoma cases, which is especially important when trying to avoid falsely diagnosing ill patients as healthy (Fränti & Mariescu-Istodor, 2023).

$$Recall = TP / (TP + FN)$$

F-1 Score: Particularly useful for dealing with imbalanced datasets (Dip et al., 2022), F-1 Score uses precision and recall to produce a harmonic average of the two. While this metric does seem harder to understand in comparison to accuracy, it is generally more useful in terms of evaluating a model's ability to classify both classes correctly.

$$F1\ Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

These metrics were employed inside all three primary sources when each examined their approach to optimizing their skin disease classification models. All metrics range from 0-1, with greater values representing greater performance. When analyzing the results, all conclusions and analysis were made looking at the metrics holistically, rather than as individual metrics. With the models outputting the scores of the various metrics representing how each model did, the researcher

organized the scores based on the dataset the models were evaluated on and by the metrics themselves.

HAM10000

To start, the researcher evaluated the models on the HAM10000 testing set, which was composed of 1,561 benign images and 236 melanoma images. The results before and after data augmentation for both models are depicted in Table 6.

	ResNet50		Random Forest	
	No Augmentations	With Augmentations	No Augmentations	With Augmentations
Accuracy	0.82	0.87	0.77	0.83
Precision	0.39	0.48	0.28	0.40
Recall	0.18	0.49	0.53	0.67
F1-Score	0.25	0.49	0.37	0.50

TABLE 6. HAM10000 Results

For both models, there was only a minor improvement in accuracy when using data augmentation, with ResNet50's accuracy improving by 0.05 and Random Forest's accuracy improving by 0.06. Comparatively, the improvements across the positive metrics for both models were generally more substantial. Without data augmentation, both ResNet50 and Random Forest displayed poor scores in both precision and recall, which quantifies the poor ability in classifying malignant images in general. These findings along with the models' accuracy score strongly reaffirm that biases were developed during training when using an imbalanced dataset, as suspected prior in the Literature Review (Mumuni & Mumuni, 2022). It was when data augmentation was applied that both models improved significantly across all positive metrics. ResNet50 obtained a precision score of 0.48, a recall score of 0.49, and an F1-score of 0.49. Likewise, Random Forest obtained a precision score of 0.40, a recall score of 0.67, and an F1-score of 0.50. Because data augmentation provided new and unique malignant images for which the models could train on as well as reduced the imbalance present in the training set, the models were able to better learn the defining features that distinguish malignant images from benign ones. Thus, data augmentation was effective in mitigating much of the effects of the class imbalance that existed originally.

Diverse Dermatology Images

After evaluating the models on the HAM10000, the researcher evaluated the models by directly applying the models trained on HAM10000 to the DDI dataset. As mentioned before, the main difference between the datasets remains in the skin color of the images as well as how the images were taken. For the DDI dataset, the images were not professionally taken nor dermoscopic, and the patients/individuals captured were of darker skin colors.

	ResNet50		**Random Forest**	
	No Augmentations	**With Augmentations**	**No Augmentations**	**With Augmentations**
Accuracy	0.44	0.55	0.43	0.44
Precision	0.10	0.17	0.11	0.15
Recall	0.33	0.52	0.42	0.57
F1-Score	0.15	0.26	0.17	0.24

TABLE 7. DDI Results

Overall, both models struggled with classifying the darker-skinned images of the DDI dataset, as seen by the dip in all metrics apart from recall. Once again, this is likely due to the large differences just mentioned between the images of the HAM10000 and DDI datasets.

When investigating the effectiveness of data augmentation for melanoma detection in the DDI dataset by comparing the models and their scores before and after augmentation was applied, the findings were quite similar to the results of the HAM10000 dataset. Both models saw improvements across all metrics after they were trained with the additional augmented images, especially in the positive metrics. For example, both models saw the greatest increase in performance in their recall scores. However, only minor improvements in the precision and accuracy scores signifies that data augmentation is slightly hindering the models' ability at classifying healthy images as well. Both trends signal back to the idea that data augmentation also affects the frequency or bias for which a model has toward classifying images as a specific label. With more malignant images in their training dataset, it is likely that the models are becoming accustomed to making more malignant classifications, significantly increasing the amount of correct malignant classifications (major improvement in recall) while also increasing the amount of incorrect malignant classifications (minor improvement in accuracy and precision).

Discussion

This study aimed to examine the effectiveness of data augmentation for improving machine learning models' performance in diagnosing

melanoma images in darker and more diverse skin ranges. Prior studies using machine learning to pursue the problem of accurate image classification of other diseases have demonstrated the benefits and advantages of using data augmentation to improve the overall quality of training data, leading to enhanced model performances. Likewise, while the overall performances of the proposed models were generally mediocre, comparing the performances before and after data augmentation showed a slight increase across most metrics when evaluating the models on the HAM10000 dataset, reaffirming the findings of previous studies, such as the study conducted by Nawaz and his colleagues' (Nawaz et al., 2025). Such was similarly seen when testing the models on the diverse images of the DDI, although to a significantly lesser extent, exemplifying the challenge of developing well-rounded and inclusive models with the insufficient amount of diverse data that is currently available. Concerning how the performance metrics produced by this study stand against the current state of medical diagnostic models developed to tackle the issue of melanoma detection in darker skin groups, this study was not able to eclipse the accuracy and positive metric scores of Dip et al. and their approach of directly using the images of the DDI dataset to finetune previous models.

Limitations

There were many obstacles and restraints to this research study that must be addressed. One of the major limitations was access to sufficient computational power required for intensive model training. Training machine learning models is a highly computationally consuming task that requires large quantities of memory and graphics processing units (Yazici et al., 2023). The researcher primarily relied on their home computer, which is a HP All-in-One 22 computer with only 8 gigabytes of memory and AMD Radeon Vega 3 Graphics chip. Therefore, many models, such as the Support Vector Machine algorithm (SVM) were unable to be included into the scope of this study due to the inability to intensively train these models. Furthermore, this limited computational power restricted the amount of augmentations the researcher was able to apply to a singular image, meaning that additional images containing several different augmentations could not be added. Additionally, while this study acknowledges the lack of substantial diverse datasets, with the size of the DDI testing set being less than 150 images, and with only 21 of them representing the malignant class, the real performance of the models was still difficult to gauge.

Conclusion

This research assessed the effectiveness of data augmentation in improving the training and accuracy of two machine learning algorithms for melanoma skin cancer detection in diverse skin groups. One of these models was the state-of-the-art ResNet50 neural network, while the other

was the Random Forest ensemble machine learning algorithm. The models were evaluated once using a training set without augmented images characterized by a large class imbalance between the malignant and benign image classes. These metrics were then compared to the models when evaluated on a training set that used data augmentation to mitigate the aforementioned dataset imbalance. The results showed that for melanoma detection in whiter skin, both ResNet50 (F1-score = 0.49) and Random Forest (F1-score = 0.50) were able to detect more of the melanoma images with the assistance of the additional augmented images. The advantages of data augmentation, however, were less present in the evaluation of models against the diverse testing dataset composed of darker skin groups, in which ResNet50 and Random Forest achieved marginal improvements across a majority of the metrics used.

Despite the subpar results, this study still holds value for medical image classification applications. Notably, the findings of this study further cement the use of data augmentation to improve model generalization by overcoming class imbalances that have been established by prior studies (Nawaz et al., 2025), especially for underrepresented groups such as the Black and Hispanic populations. Additionally, this study underscores the severity for which the scarcity of diverse skin cancer datasets has hampered the proper development of accurate models that can generalize to any racial group. In future work, models that were not included in this study (i.e. Support Vector Machine) should be examined in tandem with data augmentation to expand the amount of unique models tested. Additionally, future research should also experiment with the use of several other augmentation techniques (i.e. changes in image brightness) to help with synthesizing diverse training data.

References

Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. Med, 2(6), 642–665. https://doi.org/10.1016/j.medj.2021.04.006

Aggarwal, P., Mishra, N. K., Fatimah, B., Singh, P., Gupta, A., & Joshi, S. D. (2022). Covid-19 image classification using Deep Learning: Advances, challenges and opportunities. Computers in Biology and Medicine, 144, 105350. https://doi.org/10.1016/j.compbiomed.2022.105350

Ajayi, O. G., & Ashi, J. (2023). Effect of varying training epochs of a faster region-based convolutional neural network on the accuracy of an automatic weed classification scheme. Smart Agricultural Technology, 3, 100128. https://doi.org/10.1016/j.atech.2022.100128

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M.

(2023). Revolutionizing Healthcare: The role of Artificial Intelligence in Clinical Practice – BMC Medical Education. BioMed Central. https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-023-04698-z

Alrabai, A., Echtioui, A., & Kallel, F. (2025). Exploring pre-trained models for skin cancer classification. Applied System Innovation, 8(2), 35. https://doi.org/10.3390/asi8020035

American Cancer Society. (2024). Cancer Facts & Figures. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2024/2024-cancer-facts-and-figures-acs.pdf

Babar, M., Butt, R. T., Batool, H., Asghar, M. A., Majeed, A. R., & Khan, M. J. (2021). A refined approach for classification and detection of melanoma skin cancer using Deep Neural Network. International Conference on Digital Futures and Transformative Technologies (ICoDT2), 1–6. https://doi.org/10.1109/icodt252288.2021.9441520

Berrar, D. (2019). Cross-validation. Encyclopedia of Bioinformatics and Computational Biology, 542–545. https://doi.org/10.1016/b978-0-12-809633-8.20349-x

Bhatt, H., Shah, V., Shah, K., Shah, R., & Shah, M. (2023). State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: A comprehensive review. Intelligent Medicine, 3(3), 180–190. https://doi.org/10.1016/j.imed.2022.08.004

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, Z., Cen, J., & Xiong, J. (2020). Rolling bearing fault diagnosis using time-frequency analysis and deep transfer convolutional neural network. IEEE Access, 8, 150248–150261. https://doi.org/10.1109/access.2020.3016888

Choudhary, S., & Kesswani, N. (2020). Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using Deep Learning in IOT. Procedia Computer Science, 167, 1561–1573. https://doi.org/10.1016/j.procs.2020.03.367

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. PLOS Digital Health, 3(11). https://doi.org/10.1371/journal.pdig.0000651

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

clinical image set. Science Advances, 8(32). https://doi.org/10.1126/sciadv.abq6147

Danquah, R. A. (2020). Handling imbalanced data: A case study for binary class problems. arXiv.org. https://arxiv.org/abs/2010.04326

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2009.5206848

Dip, S. A., Arif, K. H. I., Shuvo, U. A., Khan, I. A., & Meng, N. (2024). Equitable skin disease prediction using transfer learning and domain adaptation. arXiv.org. https://arxiv.org/abs/2409.00873

Fränti, P., & Mariescu-Istodor, R. (2023). Soft precision and recall. Pattern Recognition Letters, 167, 115–121. https://doi.org/10.1016/j.patrec.2023.02.005

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. Information and Software Technology, 162, 107268. https://doi.org/10.1016/j.infsof.2023.107268

Goz, E., Yuceer, M., & Karadurmus, E. (2019). Total organic carbon prediction with Artificial Intelligence Techniques. Computer Aided Chemical Engineering, 889–894. https://doi.org/10.1016/b978-0-12-818634-3.50149-1

Hassan, H., Ren, Z., Zhao, H., Huang, S., Li, D., Xiang, S., Kang, Y., Chen, S., & Huang, B. (2022). Review and classification of AI-enabled COVID-19 CT imaging models based on Computer Vision Tasks. Computers in Biology and Medicine, 141, 105123. https://doi.org/10.1016/j.compbiomed.2021.105123

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). Deep residual learning for image recognition. arXiv.org. https://arxiv.org/abs/1512.03385

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 06(02), 107–116. https://doi.org/10.1142/s0218488598000094

Jumah, F., Raju, B., Nagaraj, A., Shinde, R., Lescott, C., Sun, H., Gupta, G., & Nanda, A. (2022). Uncharted waters of machine and deep learning for surgical phase recognition in neurosurgery. World Neurosurgery, 160, 4–12. https://doi.org/10.1016/j.wneu.2022.01.020

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of Ensemble Learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert

Systems with Applications, 244, 122778. https://doi.org/10.1016/j.eswa.2023.122778

Li, H., Rajbahadur, G. K., Lin, D., Bezemer, C.-P., Ming, Z., & Jiang. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. arXiv.org. https://arxiv.org/abs/2401.10359

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine learning algorithm validation. Neuroimaging Clinics of North America, 30(4), 433–445. https://doi.org/10.1016/j.nic.2020.08.004

Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and Mitigation Strategies. Npj Digital Medicine, 6(1). https://doi.org/10.1038/s41746-023-00858-z

Morid, M. A., Borjali, A., & Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. Computers in Biology and Medicine, 128, 104115. https://doi.org/10.1016/j.compbiomed.2020.104115

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. Array, 16, 100258. https://doi.org/10.1016/j.array.2022.100258

Nawaz, K., Zanib, A., Shabir, I., Li, J., Wang, Y., Mahmood, T., & Rehman, A. (2025). Skin cancer detection using dermoscopic images with convolutional neural network. Nature News. https://www.nature.com/articles/s41598-025-91446-6

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Patel, M. J., Khalaf, A., & Aizenstein, H. J. (2016). Studying depression using imaging and Machine Learning Methods. NeuroImage: Clinical, 10, 115–123. https://doi.org/10.1016/j.nicl.2015.11.003

Patel, R. H., Foltz, E. A., Witkowski, A., & Ludzik, J. (2023). Analysis of artificial intelligence-based approaches applied to non-invasive imaging for early detection of melanoma: A systematic review. Cancers. https://pmc.ncbi.nlm.nih.gov/articles/PMC10571810/

Ravi, V. (2022). Attention cost-sensitive deep learning-based approach for skin cancer detection and classification. Cancers, 14(23), 5872. https://doi.org/10.3390/cancers14235872

Salzberg, S. L. (1998). A tutorial introduction to computation for Biologists. New Comprehensive Biochemistry, 11–27. https://doi.org/10.1016/s0167-7306(08)60459-7

Saputra, A., & Suharjito. (2019). Fraud detection using machine learning in e-commerce. International Journal of Advanced Computer Science and Applications, 10(9). https://doi.org/10.14569/ijacsa.2019.0100943

Silva, C. N. (2010). Ex Post Facto Study. Encyclopedia of Research Design. https://doi.org/10.4135/9781412961288.n145

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. Machine Learning and the Internet of Medical Things in Healthcare, 89–111. https://doi.org/10.1016/b978-0-12-821229-5.00003-3

Thurnhofer-Hemsi, K., & Domínguez, E. (2020). A convolutional neural network framework for accurate skin cancer detection. Neural Processing Letters, 53(5), 3073–3093. https://doi.org/10.1007/s11063-020-10364-y

Tong, K., Wu, Y., & Zhou, F. (2020). Recent advances in small object detection based on Deep Learning: A Review. Image and Vision Computing, 97, 103910. https://doi.org/10.1016/j.imavis.2020.103910

Tschandl, P., Rosendahl, C., & Kittler, H. (2018, August 14). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Nature News. https://www.nature.com/articles/sdata2018161

Wah, T. N., & Htwe, H. N. (2020). Analysis of Histogram-based Image Processing Techniques with Ultimate Concepts and Idea for Advanced Studies. Engineering Technology and Scientific Journal, 2(2). https://www.researchgate.net/publication/349683278

Wainer, J. (2024). An empirical evaluation of imbalanced data strategies from a practitioner's point of view. Expert Systems with Applications, 256, 124863. https://doi.org/10.1016/j.eswa.2024.124863

Yazici, İ., Shayea, I., & Din, J. (2023). A survey of applications of Artificial Intelligence and machine learning in future mobile networks-enabled systems. Engineering Science and Technology, an International Journal, 44, 101455. https://doi.org/10.1016/j.jestch.2023.101455

Zhang, X., Zhai, D., Li, T., Zhou, Y., & Lin, Y. (2023). Image inpainting based on Deep Learning: A Review. Information Fusion, 90, 74–94. https://doi.org/10.1016/j.inffus.2022.08.033