Evaluating Gender Bias and Fairness in Skin Lesion Diagnoses using Convolutional Neural Networks

Aakash Kondaka

Received May 03, 2024
Accepted July 22, 2024
Electronic access August 31, 2024

Integrating Artificial Intelligence (AI) in medical diagnostics has shown great potential in improving the speed and efficiency of disease detection, especially in skin lesion diagnosis. This study focuses on evaluating the performance and identifying potential gender biases in AI-powered diagnosis of skin lesions using Convolutional Neural Networks (CNNs), specifically DenseNet-121 and ResNet-50. We utilized the comprehensive HAM10000 dataset, which includes diverse skin lesion types and patient demographics, to conduct a comparative analysis of the diagnostic accuracies of the models and investigate underlying gender biases. Our methodology included data preprocessing, model training focusing on feature extraction, optimization, and an exhaustive evaluation based on percent accuracy, F1 score, and Matthew's correlation coefficients. The findings reveal statistically significant differences in diagnostic performance between genders, with DenseNet-121 outperforming ResNet-50 across most metrics for both male and female datasets. We also examined the impact of data augmentation on model performance, which demonstrated significant improvements in accuracy and reliability, especially for female skin lesions. Additionally, we explored the variance in male versus female percent accuracy across different cell lesion types, revealing statistically significant differences in diagnostic performance. This study highlights the critical challenge of biases in medical AI diagnostics and proposes a path forward through diversified training datasets, algorithmic fairness enhancements, and the necessity of adopting comprehensive evaluation metrics to ensure equitable healthcare outcomes. By addressing these biases, we aim to contribute to the ethical advancement of AI in healthcare, emphasizing fairness, transparency, and inclusivity in AI-driven diagnostics.

Introduction

Skin cancer, characterized by the uncontrolled growth of abnormal skin cells, has emerged as one of the most common forms of cancer worldwide. Approximately 5 million new cases of skin cancer are reported in the United States each year. Its most lethal variant, melanoma, only accounts for 1% of skin cancer cases yet is the cause for a large majority of skin cancer deaths. According to the Skin Cancer Organization, in the year 2024, an estimated 11.3% of skin cancer deaths will be melanoma cases. This demands urgent attention due to its rapidly increasing incidence and high potential for fatality if not detected early[1].Early and accurate diagnosis is pivotal, as it can significantly increase survival rates, which may exceed 90% when melanoma is identified at an initial stage. However, the visual evaluation of skin lesions, a critical step in early detection, can be challenging due to the subtle nuances of dermoscopic images, leading to potential misdiagnoses[2–4].Early detection and assessment of skin lesions are crucial for effective treatment. Still, the limited number of dermatologists compared to the number of patients is a significant issue in many developing countries. In this context, Artificial Intelligence (AI) has emerged as a transformative force in medical imaging diagnostics[5]. The deployment of Convolutional Neural Networks (CNNs) has brought about unprecedented advancements in the accuracy and efficiency of analyzing complex visual data, such as that found in dermoscopic images of skin lesions[6].Previously CNNs have been applied to various medical image diagnoses and showed promise[7–9]These AI-driven models have the potential not only to support clinicians in making more informed decisions but also to democratize healthcare by making expert-level diagnostics more widely accessible[10].

This study delves into the nuanced realm of AI-powered medical imaging, specifically addressing gender biases in diagnosing skin lesions. Gender biases in medical diagnostics can lead to significant disparities in healthcare outcomes, affecting treatment efficacy and patient prognosis. Such disparities are particularly concerning in AI-driven diagnostics, where models trained on imbalanced datasets may inadvertently learn and perpetuate existing biases. These biases can result in skewed diagnostic accuracy, where one gender may receive more accurate diagnoses than the other. This can lead to delayed or incorrect treatments, adversely affecting patient outcomes. For instance, if a model consistently underperforms in diagnosing conditions in female patients, these patients may experience delayed diagnoses and treatments, potentially worsening their prognosis. Addressing these disparities is essential to ensure

that AI models provide accurate and equitable healthcare outcomes for all patients, regardless of gender. By leveraging the comprehensive HAM10000 dataset, which features a rich variety of skin lesion types and patient demographics, we conduct an in-depth comparative analysis of two prominent CNN architectures: DenseNet-121 and ResNet-50[11,12]DenseNet-121 and ResNet-50 were selected due to their proven efficacy in medical image classification tasks and their distinct architectural advantages. DenseNet-121 is characterized by its densely connected layers that promote feature reuse across the network, enhancing its capability to capture intricate patterns in medical images. ResNet-50 utilizes residual connections to alleviate the vanishing gradient problem, enabling the training of deeper networks. Using these models, our focus is on uncovering and mitigating the gender disparities that may skew diagnostic outcomes[13].

Our objectives are manifold: to rigorously evaluate the performance of these CNN models in diagnosing skin lesions, to identify any underlying gender biases, and to propose effective strategies to mitigate these biases. We aim to enhance the fairness and accuracy of AI diagnostic tools, ensuring that the benefits of AI in healthcare are equitably distributed across all patient demographics[14].Ultimately, through a blend of detailed data analysis and the development of bias mitigation techniques, we seek to contribute to the ethical advancement of AI in healthcare, ensuring that AI not only surpasses human accuracy but also adheres to the highest standards of fairness and inclusivity.

Materials and Methods

Dataset and Preprocessing

This study utilized the publicly available HAM10000 dataset, comprising 10,015 dermatoscopic images spanning several types of skin cancer, more specifically, Actinic Keratoses, Basal Cell Carcinoma, Benign Keratosis-like Lesions, Dermatofibroma, Melanocytic Nevi, and Vascular Lesions, as well as information regarding sex. The dataset was initially curated to facilitate the training and testing of machine learning models in accurately classifying various skin lesions.

Preprocessing Steps

1. Data Cleaning: In our preprocessing steps, we removed images corresponding to unique lesions, which means excluding images that depict skin lesions not appearing elsewhere in the dataset. This was done to ensure that the dataset consists of multiple images of the same lesion type for each patient, facilitating more robust training and evaluation of the AI models. By removing these unique lesions, we aim to reduce the variability and potential noise in the dataset, thereby improving the model's ability to generalize across different lesion types.

2. Dataset Splitting: The cleaned dataset was divided into training, validation, and testing subsets following a 70:10:20 ratio, respectively. This split was designed to provide a comprehensive assessment of the model's performance across unseen data.

3. Data Augmentation: To enhance model robustness, simulate a variety of imaging conditions, and balance the dataset to counter the large class imbalances, we applied several data augmentation techniques, including random horizontal and vertical flips, rotations up to 20 degrees, and color jittering. Each image was resized to a uniform dimension of 224x224 pixels to meet the input requirements of the CNN architectures. These steps were critical in preventing overfitting and ensuring the models' generalizability to diverse imaging conditions.

4. Normalization: The images were normalized using predetermined mean and standard deviation values across the RGB channels to standardize the input data for efficient model training. Specifically, we computed the mean and standard deviation values for the dataset using the following normalization parameters: Mean : (0.49139968, 0.48215827, 0.44653124)

   Standard Deviation : (0.24703233, 0.24348505, 0.26158768) These values were calculated by normalizing the images in the training set from a pixel range of 0–255 to 0–1. The normalization process helps in centering the data around zero and scaling it to have a unit variance, which is crucial for the stability and performance of deep learning models.

Convolutional Neural Networks

In this study, Convolutional Neural Networks (CNNs) were employed for the classification of skin lesions using the PyTorch library. CNNs are a class of deep neural networks that are highly effective in processing grid-like data, such as images. They are characterized by their ability to automatically learn spatial hierarchies of features from input images, making them well-suited for tasks like image classification[15]

We employed two pre-trained CNN architectures for the classification task: DenseNet-121 and ResNet-50. These models are renowned for their efficacy in image recognition tasks, attributed to their deep, complex structures which facilitate the learning of high-level features from dermatoscopic images. While DenseNet-121 is characterized by its densely connected layers that promote feature reuse across the network, ResNet-50 utilizes residual connections to alleviate the vanishing gradient problem, enabling the training of deeper networks. ResNet-50

was specifically chosen for its balance between performance and computational efficiency. The model was fine-tuned on our dataset, with all layers participating in the training process to allow for the adaptation of pre-learned filters to the specific features of dermatoscopic imagery.

Model Training

1. Feature Extraction: We set the model parameters to allow for feature extraction, ensuring that only the final classification layers of the networks were trained on our dataset. This approach leverages the pre-learned filters of these networks, fine-tuning them to the specifics of skin lesion classification.

2. Optimization and Loss Function: The Adam optimizer was utilized with a learning rate of 1e-3, alongside a cross-entropy loss function, to handle the multi-class classification problem efficiently. This combination was chosen for its effectiveness in sparse gradient handling and its robustness in multi-class settings. Cross-entropy loss is the chosen loss function for training the neural network model. For binary classification tasks, the cross-entropy loss is a standard choice, measuring the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label, given by:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \quad (1)$$

   where N is the number of observations, y_i is the binary indicator (0 or 1) if the class label is the correct classification for observation i, and $\hat{y}_i$ is the predicted probability of observation i being of the positive class. This particular loss function evaluates the disparity between predicted class probabilities and the actual class labels. By minimizing the cross-entropy loss, the model aims to improve its accuracy in predicting the correct class labels for the given input data, ultimately enhancing its performance in the classification task.

3. Environment Setup: The training was conducted using PyTorch on a CUDA-enabled device to leverage GPU acceleration, ensuring efficient processing and model optimization.

4. Hyperparameters: Key hyperparameters included the learning rate, batch size (set to 32 for manageable memory consumption while allowing for sufficient gradient approximation), and the number of epochs (30), determined empirically to balance between underfitting and overfitting.

Evaluation metrics

True Positive, True Negative, False Negative, and False Positive. True positives (TPs) occur when the model correctly detects a lesion, directly impacting patient care by ensuring timely treatment for actual conditions. True negatives (TNs) represent instances where the model accurately identifies the absence of a lesion, avoiding unnecessary anxiety and medical procedures for the patient. False positives (FPs) occur when a model incorrectly identifies a lesion when none exists, while false negatives (FNs) happen when the model fails to detect an existing lesion. These errors are crucial in medical diagnostics where FPs can lead to unnecessary anxiety and procedures, and FNs can result in untreated medical conditions. By comprehensively analyzing TPs, TNs, FPs, and FNs in our experiments, we aim to uncover how biases might skew diagnostic accuracy across different gender categories. This analysis allows us to pinpoint specific challenges and limitations of CNN-based methods in mitigating biases and ensuring equitable healthcare AI.

Accuracy. Accuracy is the ratio of correctly predicted observations to the total observations. It is calculated as

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (2)$$

However, accuracy alone is insufficient in imbalanced datasets as it may not reflect the model's performance across different classes. High accuracy rates indicate that the model is performing well overall in terms of making correct predictions for both classes. However, a high accuracy can be misleading in the presence of class imbalances, as it might simply reflect the dominance of the majority class. By examining accuracy rates across gender categories and demographic attributes, we aim to gauge the extent to which biases impact the overall diagnostic capabilities of AI systems in healthcare. This analysis allows us to quantify disparities in model performance and identify potential areas for intervention to promote fairness and equity in healthcare AI.

F1 Score: The F1 score takes into account both precision and recall, providing a balanced assessment of a model's ability to correctly classify positive and negative instances. It is particularly useful for uneven class distributions and is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (4)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

In our research, the F1 score serves as a key indicator of model performance, allowing us to quantify the effectiveness of CNN-based methods such as DenseNet-121 and ResNet-50 in detecting skin lesions across different demographic

groups, including gender categories. By analyzing the F1 scores, we aim to gain insights into the impact of biases on model accuracy and identify areas for improvement to ensure more equitable diagnostic practices in healthcare AI. A high F1 score indicates that the model has a balanced performance between precision and recall, making it robust against class imbalances. Conversely, a low F1 score suggests that the model is lacking either in its ability to correctly identify true positives or in its conservation against false positives.

Matthew's Correlation Coefficient. Unlike simple accuracy metrics, Matthew's correlation coefficient (MCC) considers the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions, providing a more comprehensive assessment of model performance, particularly in imbalanced datasets. It is a suitable metric for evaluating models in cases of class imbalance. MCC values range from -1 to 1, with 1 indicating perfect agreement, 0 indicating random prediction, and -1 indicating complete disagreement between predicted and true values [16]. The MCC is computed as:

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (6)$$

T-Test. To statistically evaluate the differences in model performance between demographic groups, we employed t-tests. The t-test is a statistical method used to determine if there is a significant difference between the means of two groups. In our context, t-tests were used to compare the performance metrics (such as accuracy, F1 score, and MCC) of the AI models between different demographic groups (e.g., male vs. female). We chose t-tests because t-tests are straightforward to implement and interpret, making them suitable for comparing the mean performance metrics (e.g., accuracy, F1 score) between two demographic groups (e.g., male vs. female). Also, t-tests provide sufficient power to detect meaningful differences between groups, making them suitable for our analysis of performance metrics across different demographic groups. While other statistical methods, such as ANOVA or non-parametric tests, could also be used, the t-test's simplicity, robustness, and appropriateness for our data characteristics made it the preferred choice for this study. Independent sample t-tests were chosen over paired t-tests for this analysis because the comparisons were made between distinct groups of male and female patients. In this study, the same lesions were not analyzed for both genders; rather, separate groups of male and female patients with their respective lesions were considered. Therefore, the data for males and females were independent of each other. The independent sample t-test allowed us to compare the means of the two groups while accounting for the variability within each group, thus offering a robust method for evaluating gender-specific performance differences in the model.

Additionally, we considered results statistically significant if the p-value was less than 0.01. The choice of a more stringent significance level (0.01 instead of the conventional 0.05) was motivated by the need to minimize the likelihood of Type I errors, where a true null hypothesis is incorrectly rejected. In the context of our study, which involves multiple comparisons across different demographic groups and model performance metrics, a lower p-value threshold helps to ensure that the observed differences are not due to random chance. Using a p-value of 0.01 also aligns with the higher standards often required in medical research, where the implications of incorrect conclusions can be substantial. By adopting this more conservative threshold, we aim to enhance the reliability and robustness of our findings, ensuring that any detected biases or performance disparities are truly significant and not artifacts of random variation.

The t-test statistic is computed as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (7)$$

Cohen's d. Cohen's d is a measure of effect size that quantifies the difference between two means in terms of standard deviation. It is particularly useful for understanding the practical significance of differences observed between groups in various contexts, including model performance metrics in this study. The formula for Cohen's d is:

$$d = \frac{M_1 - M_2}{SD} \quad (8)$$

where M_1 and M_2 are the means of the two groups, and SD is the pooled standard deviation of the two groups.

In this study, Cohen's d was used to measure the effect sizes for differences in model performance metrics between male and female datasets, as well as between augmented and non-augmented data. The interpretation of Cohen's d follows conventional thresholds:

- Small effect size: $d \approx 0.2$
- Medium effect size: $d \approx 0.5$
- Large effect size: $d \approx 0.8$

By calculating Cohen's d, we aim to provide a clearer understanding of the practical significance of the observed differences, complementing the statistical significance indicated by p-values. This approach ensures that the reported differences are not only statistically significant but also meaningful in practical terms, enhancing the robustness of our conclusions regarding model performance and fairness.

Reproducibility

The code for dataset preparation, model training, and evaluation as well as the code I used as a frame have been shared alongside

this paper to ensure reproducibility. Specific attention was given to setting random seeds for NumPy and PyTorch to ensure consistent results across runs [17]

Background

The advent of AI in medical imaging has been transformative, offering significant strides in diagnostic capabilities. CNNs, in particular, have become instrumental in processing medical imagery due to their ability to analyze and interpret complex visual data. Despite these advances, the rapid integration of AI tools in clinical settings has unveiled a critical issue: the perpetuation of biases. Gender bias, for instance, emerges as a consequential factor influencing diagnostic accuracy, with potential ramifications for patient care and treatment outcomes. This background explores the origins and impacts of such biases, reviewing existing literature on the effectiveness of AI-driven diagnostic tools across different demographic groups. By delineating the current understanding and gaps within AI's application in medical imaging, this section sets the foundation for the presented study's aim to scrutinize and mitigate gender biases in CNN-based diagnostics [18].

Fairness and Biases in AI

Recently, the rapid adoption and increasing reliance on AI-driven diagnostic tools in healthcare have brought to light significant concerns regarding biases inherent in these technologies. Biases in AI systems, particularly in medical imaging, can arise from many sources, including but not limited to the data on which models are trained, the algorithms themselves, and the interpretative frameworks used to understand their outputs [19].These biases, if unchecked, have the potential to perpetuate and amplify existing inequalities in healthcare delivery, leading to differential diagnosis accuracy, treatment efficacy, and patient outcomes across diverse populations. Roselli, Matthews, and Talagala (2019) address the pervasive issue of bias in AI systems, emphasizing the need for systematic approaches to identify, quantify, and mitigate biases. Their work proposes a comprehensive framework to manage bias across three key areas: the transition from business intent to AI implementation, the distribution of training data, and the biases inherent in individual input samples. This robust framework includes substantiating assumptions, vetting training data, evaluating for bias, and monitoring production data, underscoring the necessity of a multidimensional approach to ensure the fairness and transparency of AI systems [20].

Fairness and Biases in Medical Imaging

Gender bias, a specific category of bias, poses a critical challenge in ensuring the equitable application of AI in medical imaging. Despite advancements in the field, studies have indicated that AI-driven diagnostic tools can exhibit varying levels of performance across different gender groups, largely due to disparities in gender representation within training datasets and underlying algorithmic assumptions, such as the homogeneity of training data and the equal representation of all demographic groups, can contribute to gender bias. These assumptions may lead to models that perform well on average but poorly on underrepresented groups. For instance, if the training data is not balanced across genders, the model may learn features that are more representative of the majority group, thereby reducing its diagnostic accuracy for the minority group. Furthermore, certain architectural choices, like the selection of specific loss functions or optimization techniques, may implicitly favor patterns that are more prevalent in one gender over another, exacerbating the bias. Zong, Yang, and Hospedales (2022) introduced MEDFAIR, a novel framework designed for benchmarking fairness in medical imaging AI models. MEDFAIR operates by systematically evaluating AI models on multiple dimensions of fairness, including demographic parity, equalized odds, and disparate impact. The framework employs a comprehensive benchmarking suite that includes both statistical and algorithmic fairness metrics to identify and quantify biases in AI models. MEDFAIR's multi-faceted approach to fairness evaluation not only assesses the performance of AI models across different demographic groups but also incorporates fairness-aware training algorithms to mitigate identified biases. The framework integrates techniques such as reweighting, which adjusts the importance of different samples during training, and adversarial debiasing, where models are trained to be indistinguishable in their performance across demographic groups. MEDFAIR also provides a detailed breakdown of fairness metrics, offering insights into how biases manifest in different stages of the AI pipeline, from data collection to model deployment. By examining biases across various AI models and datasets, the MEDFAIR framework emerges as a pivotal tool for advancing the discourse on fairness in AI, particularly within the healthcare domain. The study expands upon how the MEDFAIR framework systematically quantifies biases in AI models by evaluating performance disparities across different demographic groups, including gender. It also highlights the significant impact of model selection strategies on fairness outcomes. It was found that certain strategies, such as minimax Pareto selection, can improve the performance for the worst-case group without substantially reducing overall accuracy, thereby promoting gender fairness. This helps in identifying specific areas where gender bias is prevalent [21].

Additionally, Drukker et al. (2023) embarked on a comprehensive examination of bias across the AI development pipeline in medical imaging, emphasizing the need for strategies to ensure fairness. Through a multi-institutional analysis, the paper categorizes potential biases and offers targeted mitigation

	DenseNet-121			ResNet-50		
	% Accuracy	F1 Score	MCC	% Accuracy	F1 Score	MCC
Male	92.74% (±0.79%)	0.63	0.67	89.80% (±0.84%)	0.57	0.60
Female	88.18% (±0.52%)	0.65	0.74	84.81% (±0.68%)	0.60	0.68

Table 1: Performance Metrics of DenseNet-121 vs. ResNet-50 in Skin Lesion Diagnosis

strategies, making a significant contribution to the ongoing effort to address gender biases in AI-driven diagnostics. This work provides a critical framework for understanding and mitigating bias in medical imaging AI, emphasizing the importance of diverse and representative datasets for improving diagnostic accuracy and equity[22].

Impact of Gender Bias on Healthcare Outcomes

Gender bias in diagnostic algorithms can have profound effects on healthcare outcomes. When AI models are trained on imbalanced datasets, they may perform suboptimally for underrepresented groups, leading to diagnostic inaccuracies. For example, if a model is less accurate in diagnosing skin lesions in females, it may result in delayed diagnoses for female patients, causing the disease to progress to more advanced stages before treatment begins. This delay can significantly impact survival rates, especially for aggressive conditions like melanoma. Furthermore, misdiagnoses can lead to inappropriate treatments, which not only fail to address the actual condition but may also cause adverse side effects, compounding the patient's health issues.

Results

DenseNet-121 vs. ResNet-50 Performance

The results (Table 1) suggest a significant difference in all of the metrics. In all of our comparative analyses, we employed t-tests to rigorously evaluate the differences in performance metrics between these models. The t-test, a fundamental statistical tool, helps in assessing whether the means of two groups are statistically different from each other. Generally, a p-value (p) less than or equal to the significance level of 0.01 is considered statistically significant, leading to the rejection of the null hypothesis which in this case assumes there is no significant difference in diagnostic accuracy between DenseNet-121 and ResNet-50 (null hypothesis serves as a baseline assumption that there is no significant difference between the two things being compared).[16]. For male percent accuracy, DenseNet-121 outperformed ResNet-50, with a mean accuracy of 92.74% compared to 89.80% ($p<0.00001$; $t = 11.46$). As shown in Figure 1, the performance metrics indicate a variation in model accuracy between genders (% Accuracy is out of 1.00 rather than 100% in Figure 1. While accuracy is typically reported as a percentage, in this study, we report it as a fraction out of 1.00. This approach ensures consistency with other performance metrics such as MCC and F1 scores, which are also scaled between 0 and 1.). This suggests that DenseNet-121's architecture might be more adept at recognizing patterns relevant to male skin lesions. The standard deviation was marginally smaller for DenseNet-121 (0.79%) than ResNet-50 (0.84%), indicating slightly more consistency across trials for DenseNet-121. This lower standard deviation suggests that DenseNet-121's performance is more stable and reliable compared to ResNet-50. In practical terms, a lower standard deviation means that DenseNet-121 is less sensitive to variations in the input data and can consistently produce accurate results across different subsets of the dataset. This consistency is crucial in medical diagnostics, where reliability and reproducibility of results are paramount. High variability, on the other hand, could indicate that a model's performance is heavily influenced by specific data characteristics, which might not generalize well to new, unseen data. The slightly higher standard deviation for ResNet-50 indicates that its accuracy is more variable, potentially making it less reliable in clinical settings. This variability might be due to the model's architecture, which could be more prone to overfitting or underfitting certain types of skin lesions, leading to fluctuating performance metrics. Therefore, the lower standard deviation for DenseNet-121 not only highlights its robustness but also underscores its suitability for real-world applications where consistent and dependable performance is critical.

When considering the F1 scores for males, DenseNet-121 exhibited a higher mean score (0.6277) relative to ResNet-50 (0.5748), with the difference being statistically significant ($p<0.00001$; $t = 7.578$). The standard deviations for both models were comparable, suggesting that the observed differences in mean scores are substantive and not merely a reflection of variability in the models' performances. Moreover, the analysis of MCC, which is considered a balanced measure of binary classification performance, revealed that DenseNet-121 had a higher mean MCC for males (0.6651) than ResNet-50 (0.6025), with a t-statistic of 8.806, confirming the significant difference

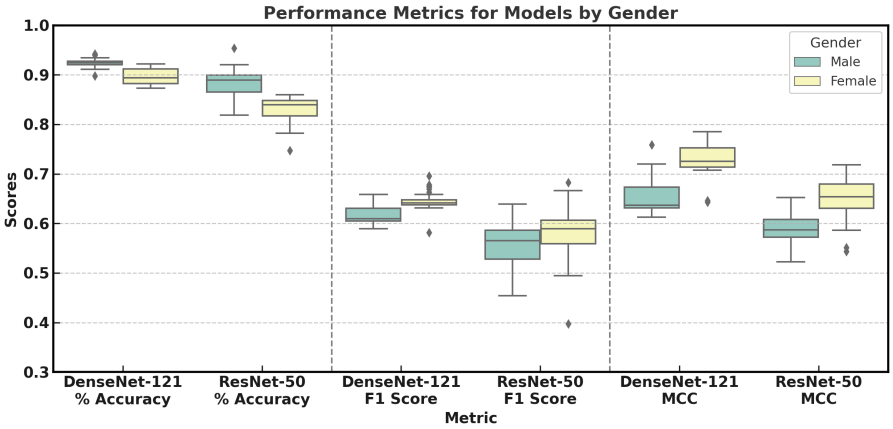

Fig. 1 Comparative Performance Metrics by Gender for DenseNet-121 and ResNet-50 Models.

(p<0.00001). This finding corroborates the results obtained for accuracy and F1 score, indicating a consistently superior performance by DenseNet-121 in diagnosing male skin lesions. For females, DenseNet-121 again outpaced ResNet-50 with a mean accuracy of 88.18% against 84.81% ($p < 0.00001$; t = 17.567). The heightened t-statistic here in comparison to male accuracy suggests an even more pronounced advantage of DenseNet-121 over ResNet-50 for female datasets. This again suggests that DenseNet-121's architecture might be more adept at recognizing patterns relevant to female skin lesions. The standard deviation was higher for ResNet-50 (0.68%), implying greater variability in performance across trials for this model. The F1 scores for females paralleled the pattern observed in males, with DenseNet-121 attaining a higher mean score (0.6521) than ResNet-50 (0.6010), again confirming a statistically significant difference (p<0.00001; t = 6.642). This result indicates that DenseNet-121 maintains a more balanced precision and recall in diagnosing female skin lesions. Lastly, the MCC for females also favored DenseNet-121 (mean of 0.7415) over ResNet-50 (mean of 0.6755), with a statistically significant t-statistic of 7.213 ($p < 0.00001$). This again suggests that DenseNet-121 maintains a more reliable performance in its binary classification of skin lesions in females just like results for F1 scores and % Accuracy suggest. The higher performance metrics for females in both models challenge traditional expectations, as accuracy percentages were higher for males. These results may imply that while the models are more frequently correct in their overall predictions for males, they demonstrate a finer discrimination between lesion classes in females. The reasons behind this discrepancy may include intrinsic differences in skin lesions between genders or biases within the models themselves, necessitating further inquiry. Several other factors may contribute to this performance disparity. First, if the dataset includes a well-balanced representation of female skin lesions, the model can learn more effectively from these examples, leading to better discrimination between lesion classes. The higher MCC and F1 scores for females indicate that the model has learned to identify true positives and minimize false classifications more effectively for females. Second, DenseNet-121's architecture, with its densely connected layers, allows it to capture more intricate and relevant features from the dermatoscopic images. These features may be particularly informative for distinguishing between different types of lesions in females. Third, the intrinsic differences in skin lesions between genders might mean that the features used by the model are more distinct and easier to learn for females, resulting in better performance metrics for females even if the overall accuracy is higher for males. Finally, bias mitigation techniques such as data augmentation and reweighting of underrepresented classes might have been more effective for female patients, leading to improved performance in terms of MCC and F1 scores.

Additionally, to further understand the practical significance of these differences, we calculated the effect sizes for the observed differences in diagnostic accuracy between male and female datasets for both DenseNet-121 and ResNet-50. For the male data between DenseNet-121 and ResNet-50, the effect size (Cohen's d) was 3.57 indicating a large effect. Similarly for females Cohen's d = 5.47 again indicating a large effect. This suggests that DenseNet-121 performs significantly better than ResNet-50 in diagnosing male and female skin lesions, which has important implications for clinical practice and patient outcomes. These findings highlight not only statistically significant differences but also practically meaningful differences in model performance within each gender group. The large effect sizes underscore the importance of model selection in achieving equitable diagnostic performance across different demographics, suggesting that

DenseNet-121 should be preferred over ResNet-50 for its superior accuracy. We also conducted a detailed statistical analysis to evaluate the gender-based differences in model performance for in both the DenseNet-121 and ResNet-50 model. For DenseNet-121, the mean accuracy for males was 92.74% ($\pm$0.79%), while for females, it was 88.18% ($\pm$0.52%). The calculated effect size (Cohen's d) was 6.32, indicating a substantial practical difference in diagnostic accuracy between male and female datasets. The t-test results (t-statistic = 21.57, $p < 0.00001$) confirm that this difference is statistically significant, suggesting that the model performs significantly better on the male dataset compared to the female dataset. Similarly, for ResNet-50, the mean accuracy for males was 89.80% ($\pm$0.84%), and for females, it was 84.81% ($\pm$0.68%). The effect size (Cohen's d) was 5.64, also indicating a significant practical difference in diagnostic accuracy between male and female datasets. The t-test results (t-statistic = 20.65, $p < 0.00001$) further confirm the statistical significance of this difference, with the model showing better performance on the male dataset. These findings highlight not only statistically significant differences but also practically meaningful differences in model performance across genders. The large effect sizes underscore the importance of addressing potential biases in training datasets and model architectures to ensure equitable diagnostic performance for both genders. The superior performance of DenseNet-121 across all measures and for both genders indicates that its architectural features are particularly suited for the task at hand. The consistency of DenseNet-121's performance, as indicated by the lower standard deviations in accuracy and the substantial t-statistics across all tests, underscores its robustness as a diagnostic tool in this medical imaging context. In conclusion, while both DenseNet-121 and ResNet-50 are competent CNNs for the task of skin lesion classification, DenseNet-121 demonstrates a statistically significant superiority. Notably, the divergence between standard accuracy and MCC/F1 scores calls for a multifaceted approach to performance evaluation, ensuring a more comprehensive understanding of model efficacy and potential biases. This analysis underscores the imperative for diversified datasets and underscores the necessity of multi-metric evaluation in the development and assessment of CNNs for medical diagnostics.

Augmentation vs Non-Augmentation Impact on CNN Performance

The study further investigated the impact of data augmentation on the diagnostic performance of two CNN architectures, DenseNet-121 and ResNet-50, in skin lesion classification. Data augmentation is a widely adopted technique in machine learning to enhance the diversity and volume of training data, which can potentially lead to improved model generalizability and robustness. In the implementation of our study, we augmented the training images to enhance the model's ability to generalize across diverse conditions, thereby potentially increasing its diagnostic accuracy. The training dataset, initially comprising 8,360 images, expanded to approximately 34,345 images with augmentation. Specifically, the augmentation pipeline included resizing all images to match the CNN input requirements, applying random horizontal and vertical flips, random rotations up to 20 degrees, and subtle variations in brightness, contrast, and hue. This approach was designed to simulate a wider array of skin lesion appearances and imaging conditions, thereby training our models, DenseNet-121 and ResNet-50, on a richer dataset. The validation and test sets, in contrast, were subjected only to resizing and normalization to maintain the integrity of real-world data representation.

As shown in Figure 2, for DenseNet-121, augmentation appears to have had a mixed impact on performance metrics. Male percent accuracy showed a marginal improvement with augmentation (92.74% augmented vs. 92.15% not augmented, $p = 0.0145$, $t = 2.56$. Null hypothesis assumes no significant difference in diagnostic accuracy between CNN architectures trained with augmented and non-augmented training sets), indicating that the additional variety in training data might have slightly enhanced the model's ability to generalize. Similarly, female percent accuracy significantly benefited from augmentation (88.18% augmented vs. 91.14% not augmented, $p < 0.00001$, $t = 2.617$), suggesting that the diversity in augmented data is particularly advantageous in improving the model's performance for female samples.

Looking at the F1 scores, the augmented data led to a statistically significant improvement for both males ($p < 0.00001, t = 4.079$) and females ($p = 0.0016, t = 2.521$), albeit the absolute differences were moderate. The improvement in the F1 score, which balances precision and recall, is critical as it implies that augmentation aids the model in making more accurate positive predictions while reducing false positives and negatives.

The most pronounced impact of augmentation in DenseNet-121 was observed in the MCC, especially for females (0.7415 augmented vs. 0.71907 not augmented, $p < 0.00001$, $t = 13.705$). Since MCC is a robust metric accounting for all categories of the confusion matrix, this substantial increase indicates that augmentation significantly enhances the model's predictive quality and reliability.

The results for ResNet-50 (Figure 3) show the influence of augmentation is evident and pronounced across all metrics. Male percent accuracy improved notably with augmentation (89.79% augmented vs. 86.53% not augmented, $p < 0.00001, t = 4.763$). This significant increment underscores the effectiveness of augmentation in addressing overfitting and enhancing the model's ability to generalize from the training data to unseen data.

The F1 score for males also saw a significant increase ($p <$

0.00001, $t = 3.421$), reinforcing the benefit of augmentation in achieving a balanced precision-recall trade-off. Similarly, for females, both percent accuracy ($p < 0.00001, t = 6.643$) and F1 scores ($p < 0.00001, t = 2.966$) were significantly higher with augmentation, highlighting the augmentation's role in improving model sensitivity and specificity for both genders.

MCC for ResNet-50, like DenseNet-121, displayed improvements with augmentation for both genders ($p < 0.00001$ for males and females), confirming that the augmentation process enhances the overall classification correctness of the model.

Again, to further understand the practical significance of these differences, we calculated the effect sizes for the observed differences in diagnostic accuracy between augmented and non-augmented data for both DenseNet-121 and ResNet-50. For DenseNet-121 male accuracy, the effect size (Cohen's d) between augmented and non-augmented data is 0.71, indicating a medium effect size. This suggests that data augmentation provides a noticeable improvement in model performance, enhancing the model's ability to generalize better to unseen data. For females, the Cohen's d value is 4.44, which is a large effect size. This indicates a substantial practical difference in diagnostic accuracy between augmented and non-augmented data, highlighting the significant impact of data augmentation in improving the model's performance performance for female skin lesions, suggesting that augmentation helps the model capture more diverse and relevant features. For ResNet-50 the results were similar with the Cohen's d for male accuracy being 3.12 and 3.06 for females again indicating that there were substantial practical differences in diagnostic accuracy between augmented and non-augmented data. The large effect sizes observed suggest that augmentation techniques should be consistently applied to training datasets to ensure models can generalize well and provide accurate diagnoses across diverse patient demographics.

Augmentation's positive impact on both DenseNet-121 and ResNet-50 emphasizes its role in model training for complex tasks such as medical image diagnosis. By providing a more diversified dataset through augmentation, both CNN architectures not only enhanced their ability to accurately classify skin lesions but also improved their reliability, as evidenced by the increase in MCC. This result is pivotal in medical applications where the cost of misdiagnosis is high. The significant improvements across all metrics with augmented data suggest that employing such techniques is essential for training robust diagnostic models. It is especially noteworthy that augmentation has a considerable effect on the female dataset, which could be an indication of the initial training data lacking sufficient variability in representing female skin lesions. This finding advocates for the necessity of employing comprehensive augmentation strategies to ensure equitable performance across gender lines, enhancing the model's diagnostic capability and clinical applicability.

Bias Across the Different Classes

Using the DenseNet-121 model, independent sample t-tests were conducted to compare the percent accuracies between male and female participants across several cell types, each representing a distinct category of skin lesions. In Figure 4, we present examples of the skin lesion types analyzed in this study: Benign keratosis-like lesions, Dermatofibroma, Vascular lesions, Basal cell carcinoma, Melanocytic nevi, and Actinic keratoses, each demonstrating the characteristic visual features that the DenseNet-121 model was trained to recognize and classify. The aim was to investigate whether gender differences exist in the diagnostic accuracies for these skin lesion types.

As shown in Figure 5, the average percent accuracies for each gender within these cell types were observed as follows: For Actinic Keratoses, females had an average percent accuracy of 49.55% (SD = 4.64%), while males exhibited a lower average of 42.50% (SD = 13.08%). In Basal Cell Carcinoma, females achieved an average percent accuracy of 88.93% (SD = 3.65%), whereas males presented a significantly higher average of 96.25% (SD = 5.88%). Dermatofibroma saw females with an average percent accuracy of 53.45% (SD = 15.34%) compared to males, who had a higher average of 66.97% (SD = 18.93%). Melanocytic Nevi had females at an average percent accuracy of 96.12% (SD = 0.64%) and males slightly higher at 97.06% (SD = 0.45%). Vascular Lesions showed negligible difference between genders, with females at 84.16% (SD = 3.73%) and males at 84.28% (SD = 10.26%). Lastly, Benign Keratosis-like Lesions displayed females with an average percent accuracy of 63.67% (SD = 3.66%) and males at 49.06% (SD = 6.81%).

The t-test results revealed significant gender differences in percent accuracies for most cell types. For Actinic Keratoses, the t-test yielded a t-statistic of -2.270 and a p-value of 0.029, indicating significant differences (null hypothesis assumes no significant difference in diagnostic accuracy between males and females). Basal Cell Carcinoma showed a t-statistic of 4.732 with a p-value of 3.05e-05. Dermatofibroma had a t-statistic of 7.828 and a p-value of 1.92e-09, suggesting highly significant gender-based differences in diagnostic accuracies. Melanocytic Nevi presented a t-statistic of 5.308 and a p-value of 5.06e-06. Vascular Lesions exhibited a minimal difference with a t-statistic of 0.049 and a p-value of 0.961, indicating no significant gender difference. Lastly, Benign Keratosis-like Lesions showed a t-statistic of -9.002 with a p-value of 5.83e-11, indicating substantial gender-based differences.

These findings suggest that gender may play a significant role in the diagnostic accuracies of certain skin lesions, particularly in Basal Cell Carcinoma, Dermatofibroma, Melanocytic Nevi, and Benign Keratosis-like Lesions. The observed significant differences warrant further investigation into the underlying factors contributing to these disparities, which could have implications for the development of more tailored and effective

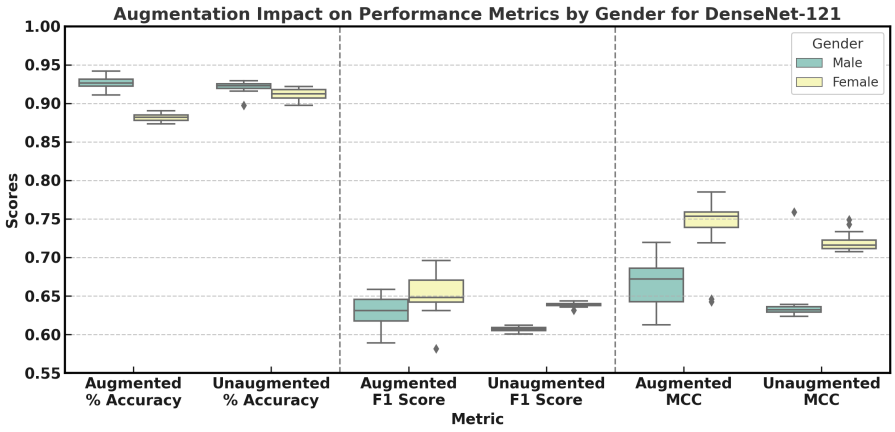

Fig. 2 Performance Comparison of Augmented vs. Unaugmented Data for DenseNet-121 by Gender

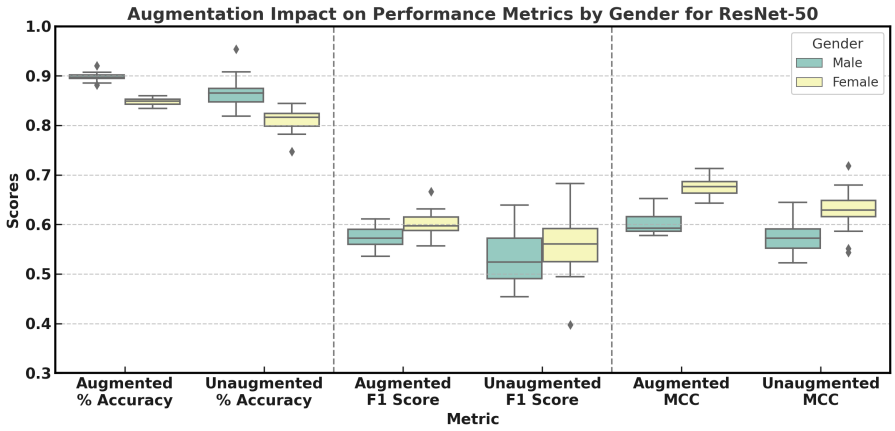

Fig. 3 Performance Comparison of Augmented vs. Unaugmented Data for ResNet-50 by Gender

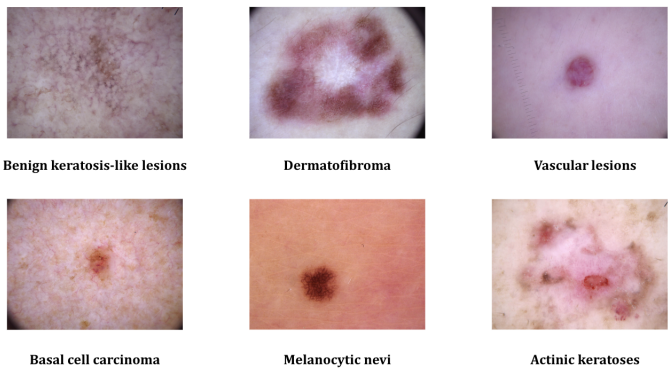

Fig. 4 Representative Dermoscopic Images of each Skin Lesions

diagnostic protocols. The average accuracies across different genders within each cell type underscore the necessity for a nuanced approach to diagnosing and treating skin lesions, taking gender as a potential factor into consideration.

Discussion

Analysis of Results

The advent of AI in medical imaging promised an era of enhanced diagnostic precision and egalitarian healthcare. The intersection of Artificial Intelligence (AI) in dermatological diagnostics, as explored through the lens of this study, exposes a delicate yet significant concern—gender biases within Convolutional Neural Network (CNN) architectures. Our findings suggest that these technologies are not immune to the prejudices that pervade human decision-making. DenseNet-121's superior performance over ResNet-50, particularly in identifying male skin lesions, prompts an inquiry into whether the architectural complexities of CNNs inherently favor features more prevalent in male skin characteristics. Such differentiation underscores the necessity for vigilance in AI training and dataset composition to mitigate gender bias. Furthermore, the augmentation techniques employed to enrich the dataset proved pivotal in amplifying the diagnostic acumen of both models, particularly for female skin lesions. This improvement highlights the importance of diverse training samples in the development of AI systems. Without adequate representation, we risk entrenching historical disparities into future technologies—antithetical to the ethos of equitable healthcare.

Additionally, while these differences are statistically significant and practically meaningful, it is crucial to explore the underlying reasons for these disparities to improve AI model performance and ensure equitable healthcare outcomes. The observed differences in model performance can be attributed to several potential factors. Biological differences in skin characteristics between genders may influence the model's performance. For instance, male and female skin may differ in terms of thickness, oil production, and susceptibility to certain types of lesions, which could affect how skin lesions are represented in dermatoscopic images. Hormonal differences can also affect skin conditions and their manifestations, potentially leading to variations in lesion appearance between genders. This might influence the model's ability to accurately diagnose lesions if it has not been trained on a sufficiently diverse dataset that accounts for these variations.

Differences in data quality could also play a significant role. The training dataset may have imbalances in the representation of male and female skin lesions, leading to biased model performance. If the dataset contains more high-quality images of lesions from one gender, the model may learn to perform better on that gender's data. Additionally, variations in image quality, lighting, and resolution can affect the model's performance. If images of female lesions are of lower quality or less consistent, this could result in poorer model performance on female data.

Model biases introduced during the training process are another critical factor. The model training process itself may introduce biases if the training data is not balanced or if the model architecture is more suited to certain types of patterns or features present predominantly in one gender's data. Algorithmic biases, such as the choice of model architecture and hyperparameters, can influence how well the model generalizes across different demographics. DenseNet-121, for instance, may have architectural features that make it more adept at recognizing certain patterns that are more common in male skin lesions, leading to better performance on male data. Additionally, several potential confounding factors might have influenced the results of our study. Age is a significant factor, as the prevalence and characteristics of skin lesions can vary with age. Older individuals, for instance, might have more pronounced or different types of lesions compared to younger individuals, which could affect the model's performance. Skin type is another crucial factor, as differences in pigmentation, texture, and other skin characteristics can influence the appearance of lesions and, consequently, the model's ability to accurately diagnose them. Lesion size is also a critical variable, as larger or more pronounced lesions might be easier for the model to diagnose accurately, while smaller or more subtle lesions could present a greater challenge.

Suggestions

To mitigate these differences and enhance the fairness and accuracy of AI models in medical diagnostics, several steps can be taken. Enhancing dataset diversity by collecting and curating more balanced datasets that include a wide variety of skin lesion images from diverse demographics is crucial. This includes ensuring adequate representation of different genders, ages, skin types, and ethnic backgrounds. Implementing fairness-aware training algorithms, such as reweighting or adversarial debiasing, can help reduce biases in model performance. These techniques can ensure that the model learns to perform equally well across different demographic groups. Regularly auditing AI models for biases and ensuring transparency in the model development and evaluation process can help identify and address potential sources of bias. This includes reporting performance metrics separately for different demographic groups. Integrating biological and clinical knowledge into the model training process can help improve the model's ability to generalize across different patient groups. This can include using domain-specific data augmentation techniques that simulate variations seen in clinical practice. To practically implement these proposed solutions, several steps can be

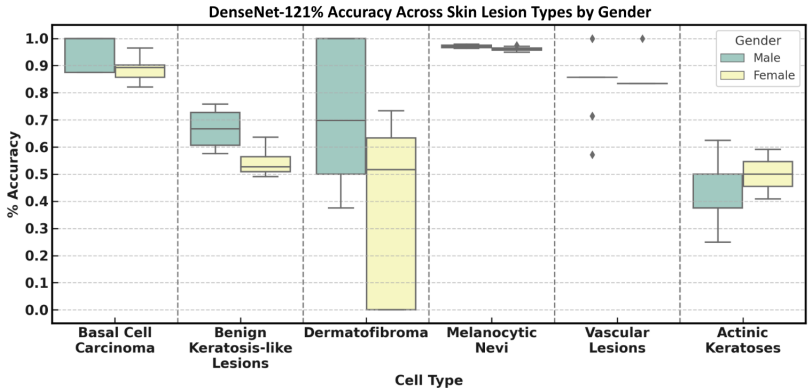

Fig. 5 Performance Comparison of % Accuracy for DenseNet-121 across Various Cell Types, Segregated by Gender

taken. First, enhancing dataset diversity requires collaboration between medical institutions, research organizations, and data repositories to gather a more representative sample of skin lesion images. This can involve initiatives to create global consortia that share de-identified patient data while ensuring ethical standards and patient privacy are maintained. Second, implementing fairness-aware training algorithms involves adapting existing machine learning frameworks to include techniques like reweighting, where underrepresented classes are given more importance during training, or adversarial debiasing, where models are trained to be indistinguishable in their performance across demographic groups. Transparent reporting and regular auditing mechanisms are essential to ensure the fairness, accuracy, and reliability of AI models in medical diagnostics. Audits should comprehensively evaluate performance metrics, focusing on accuracy, sensitivity, specificity, precision, and recall across different demographic groups to detect and quantify biases. Additionally, data quality and representation must be assessed to ensure diversity and balance, thereby reducing bias. Audits should be conducted at defined intervals: initially before deployment, periodically (quarterly or biannually), and trigger-based in response to significant changes or regulatory updates. Internal audit teams with expertise in AI, data science, and healthcare should lead these efforts, supported by independent external auditors to provide objective assessments. Multidisciplinary committees, including clinicians, data scientists, ethicists, and legal experts, should oversee the audits to ensure comprehensive evaluation. By implementing these detailed auditing mechanisms, organizations can maintain the integrity and reliability of AI systems in medical diagnostics. Additionally, integrating biological and clinical knowledge into model development can be achieved by involving domain experts in the AI development process. This collaboration can help identify relevant features and variations that should be considered during training. By exploring and addressing these potential reasons for the observed differences in model performance, we can move towards developing more equitable and accurate AI models for medical diagnostics. Ensuring that these models are free from biases is critical for providing fair and effective healthcare to all patients, regardless of gender or other demographic factors.

Examples/Existing Implementations

To practically safeguard against bias-induced errors, several integration strategies and existing implementations can be employed. One approach is the use of fairness-aware algorithms during model training. For instance, reweighting techniques adjust the importance of different training samples to ensure underrepresented groups are adequately learned by the model. Adversarial debiasing, another effective technique, involves training the model in conjunction with an adversary that attempts to predict the demographic group. The main model is then penalized for allowing the adversary to succeed, thus reducing biases. Existing implementations, such as IBM's AI Fairness 360, provide comprehensive toolkits that include these fairness-aware algorithms. These toolkits help developers evaluate and mitigate biases in their AI models. For example, IBM Watson Health employs these techniques to ensure their AI systems provide unbiased diagnostic assistance across diverse patient populations. Another practical example is Google's What-If Tool, integrated with TensorFlow, which allows users to analyze model performance across different subgroups. By using this tool, developers can visually inspect and compare model outcomes for different demographic groups, helping to identify and address potential biases during the development phase. Additionally, our research opens pathways for other future investigations, particularly in the realm of Explainable AI (XAI). XAI can play a crucial role in improving understanding and trust in AI systems used for medical diagnostics. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be used to

explain individual predictions by identifying which features most influenced the model's decision. For example, in the context of skin lesion diagnosis, these methods can highlight specific areas of an image that contributed to the classification of a lesion as malignant or benign. Techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) can generate visual explanations by creating heatmaps that highlight regions of an image that are most relevant to the model's prediction. This can help clinicians understand which parts of a skin lesion image the model focused on when making a diagnosis, providing transparency and aiding in the validation of the model's decisions. Tools like Anchor Explanations can provide high-precision rules that capture the conditions under which the model makes specific predictions. These interpretable rules can be particularly useful in clinical settings, where understanding the decision logic is critical for trust and acceptance. Developing interactive dashboards that incorporate XAI techniques can allow clinicians to interact with the AI model's predictions, explore different scenarios, and visualize explanations in real-time. These dashboards can provide a user-friendly interface for exploring model behavior and understanding the factors influencing its decisions. Incorporating regular training sessions for clinicians on how to use and interpret XAI tools can enhance their confidence and trust in AI systems. By familiarizing clinicians with the XAI methods and their outputs, they can better integrate AI into their diagnostic workflows. By implementing these XAI methods, we can enhance the transparency and trustworthiness of AI models in medical diagnostics. Ensuring that clinicians and patients understand and trust AI decisions is essential for the successful adoption of AI in healthcare. By integrating these practical strategies and leveraging existing implementations, AI systems can better safeguard against bias-induced errors, enhancing the fairness and accuracy of medical diagnostics.

Limitations

This study, while shedding light on gender biases, has its limitations. Our focus on gender might overlook other critical biases, such as those related to ethnicity, age, or skin type. Future research should expand the scope to explore these dimensions, considering the global diversity of patients. This broader focus is crucial for developing models that are not only fair and equitable but also versatile enough to adapt to diverse populations worldwide. First, the sample size used for training and evaluating the models, although substantial, may still limit the generalizability of the findings. The HAM10000 dataset includes a comprehensive collection of skin lesion images, but its size might not fully capture the diversity and complexity of skin lesions encountered in real-world clinical settings. Future research should consider larger and more diverse datasets to enhance the robustness and applicability of the models. Second, dataset diversity is a critical factor that can influence model performance. The HAM10000 dataset includes a variety of skin lesion types and patient demographics, but it may not adequately represent all demographic groups. For instance, the dataset may have imbalances in the representation of different skin types, age groups, or ethnic backgrounds. These imbalances can introduce biases into the models, affecting their accuracy and fairness. Efforts should be made to curate datasets that are more representative of the global population to ensure that AI models perform equitably across different demographic groups. Third, the generalizability of the findings is another limitation. The models were trained and evaluated on a specific dataset, and their performance may vary when applied to other datasets or in different clinical settings. The results obtained in this study may not fully generalize to other populations or healthcare environments. It is essential to validate the models on diverse datasets and in various clinical settings to ensure their broader applicability and reliability. Additionally, the study focuses on gender-based differences, but other potential sources of bias, such as ethnicity, age, or skin type, were not explicitly addressed. Future research should expand the scope to investigate these factors and their impact on model performance. By considering a broader range of potential biases, researchers can develop more comprehensive strategies for mitigating them and improving the fairness of AI models in medical diagnostics. Finally, the study's reliance on specific CNN architectures (DenseNet-121 and ResNet-50) may also limit the findings. While these models are widely used and effective, other architectures or combinations of models might yield different results. Exploring a variety of model architectures and ensemble methods could provide a more nuanced understanding of the strengths and limitations of different approaches in skin lesion diagnosis. Addressing these limitations through further research and development is crucial for advancing the field of AI in medical diagnostics. By recognizing and overcoming these challenges, we can develop more robust, fair, and generalizable AI models that contribute to equitable healthcare outcomes for all patients.

Conclusion

In the pursuit of advancing healthcare through the lens of artificial intelligence, this study has cast a spotlight on the intersection of technology and ethics, particularly in the realm of dermatological diagnostics. Our investigation into the fairness and biases within Convolutional Neural Network (CNN) architectures, using the HAM10000 dataset, has not only underscored the potential of AI to transform medical diagnostics but also highlighted the critical challenges it faces regarding gender biases. The comparative analysis of DenseNet-121 and ResNet-50 architectures revealed significant insights into the performance discrepancies across genders, with DenseNet-121 consistently outperforming ResNet-50. These findings

underscore the importance of considering algorithmic fairness in the development of AI diagnostics tools. Furthermore, the study emphasized the beneficial impact of data augmentation on improving diagnostic accuracy and reliability, particularly for female skin lesions, suggesting that diversifying training data is crucial for mitigating biases. The exploration of biases across different cell types illuminated the nuanced ways in which gender disparities manifest in AI-driven diagnostics, urging a more comprehensive approach to dataset compilation and model training. This research contributes to a growing body of work that seeks not only to harness the power of AI in healthcare but to do so ethically and equitably, ensuring that advancements in AI benefit all segments of society without perpetuating existing disparities. As we stand at the crossroads of technology and healthcare, the findings of this study serve as a call to action for the AI research community, healthcare professionals, and policymakers to collaboratively forge pathways toward the development of AI systems that are not only technically proficient but also socially just and ethically sound. By prioritizing fairness, transparency, and inclusivity, we can ensure that the AI-driven future of healthcare is one that upholds the highest standards of care for all individuals, regardless of gender or any other demographic characteristic. In conclusion, this study represents a foundational step towards reconciling the incredible potential of AI in healthcare with the imperative of ethical responsibility. It highlights the need for ongoing vigilance, innovation, and collaboration to address biases in AI-driven diagnostics, ensuring that the journey towards AI-enhanced healthcare is paved with integrity and equity. The path forward requires a concerted effort to integrate ethical considerations into the fabric of AI development, ensuring that as we advance technologically, we also progress morally, fostering a healthcare landscape that is reflective of our highest values and aspirations.

References

1 American Cancer Society, *Skin cancer*, `https://www.cancer.org/cancer/types/skin-cancer.html`, 2022.

2 Y. R. Woo, S. H. Cho, J. D. Lee and H. S. Kim, *International Journal of Molecular Sciences*, 2022, **23**, 1813.

3 R. K. Singh, R. Gorantla, S. G. R. Allada and P. Narra, *Plos One*, 2022, **17**, e0276836.

4 M. Z. Alom, T. Aspiras, T. M. Taha and V. K. Asari, *arXiv preprint arXiv:1904.11126*, 2019, **1**, 100004.

5 E. M. Senan and M. E. Jadhav, *International Journal of Computer Applications*, 2019, **178**, year.

6 P. Tschandl, N. Codella, B. N. Akay and et al., *The Lancet Oncology*, 2019.

7 D. R. Sarvamangala and R. V. Kulkarni, *Evolutionary Intelligence*, 2022, **15**, 1–22.

8 R. K. Singh and R. Gorantla, *Plos One*, 2020, **15**, e0220677.

9 R. Gorantla, R. K. Singh, R. Pandey and M. Jain, 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 397–404.

10 T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk and C. von Kalle, *Journal of Medical Internet Research*, 2018, **20**, e11936.

11 N. Hasan, Y. Bao, A. Shawon and Y. Huang, *SN Computer Science*, 2021, **2**, 389.

12 M. B. Hossain, S. H. S. Iqbal, M. M. Islam, M. N. Akhtar and I. H. Sarker, *Informatics in Medicine Unlocked*, 2022, **30**, 100916.

13 Y. Zong, Y. Yang and T. M. Hospedales, *International Journal of Computer Applications*, 2019, **178**, year.

14 T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth and B. S. Gerendas, *Nature Medicine*, 2019, **25**, 680–684.

15 MathWorks, *Convolutional neural network*, `https://in.mathworks.com/discovery/convolutional-neural-network.html`, 2020.

16 R. Gorantla, A. Kubincova, A. Y. Weiße and A. S. Mey, *Journal of Chemical Information and Modeling*, 2023.

17 X. Zhuang, *Skin lesion classification, acc: 90% [pytorch]*, `https://www.kaggle.com/code/xinruizhuang/skin-lesion-classification-acc-90-pytorch`, 2020.

18 A. King, *What do we want from fair AI in medical imaging?*, `https://www.kclmmag.org/blog_fair_ai`, 2022, Accessed: 2024-03-24.

19 S. Hooker, *Patterns*, 2021, **2**, 100241.

20 D. Roselli, J. Matthews and N. Talagala, *Software Impacts*, 2019, **1**, 100004.

21 Y. Zong, Y. Yang and T. M. Hospedales, International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020.

22 K. Drukker, W. Chen, J. Gichoya, N. Gruszauskas, J. Kalpathy-Cramer, S. Koyejo, K. Myers, R. C. Sá, B. Sahiner, H. Whitney, Z. Zhang and M. Giger, *Journal of Medical Imaging*, 2023, **10**, 061104.