MELBA Journal Machine Learning for Biomedical Imaging

On the Role of Calibration in Algorithmic Fairness Benchmarking for Skin Cancer Detection

Brandon **Dominique** [1], Prudence **Lam** [1], Nicholas **Kurtansky** [2], Jochen **Weber** [2], Kivanc **Kose** [2], Veronica **Rotemberg** [2], Jennifer **Dy** [1]

1 Northeastern University, Boston, MA, USA 02115
2 Memorial Sloan Kettering Cancer Center, New York, NY, USA 10065

stract

tificial Intelligence (AI) models have demonstrated expert-level performance in melanoma detection, yet their clinical option is hindered by performance disparities across demographic subgroups such as gender, race, and age. Previous orts to benchmark the performance of AI models have primarily focused on assessing model performance using oup fairness metrics that rely on the Area Under the Receiver Operating Characteristic curve (AUROC), which does t provide insights into a model's ability to provide accurate estimates. In line with clinical assessments, this paper dresses this gap by incorporating calibration as a complementary benchmarking metric to AUROC-based fairness trics. Calibration evaluates the alignment between predicted probabilities and observed event rates, offering deeper ights into subgroup biases. We assess the performance of the leading skin cancer detection algorithm of the ISIC 20 Challenge on the ISIC 2020 Challenge dataset and the PROVE-AI dataset, and compare it with the second- d third-place models, focusing on subgroups defined by sex, race (Fitzpatrick Skin Tone), and age. Our findings eal that while existing models enhance discriminative accuracy, they often over-diagnose risk and exhibit calibration ues when applied to new datasets. This study underscores the necessity for comprehensive model auditing strategies d extensive metadata collection to achieve equitable AI-driven healthcare solutions. All code is publicly available at tps://github.com/bdominique/testing_strong_calibration.

ywords

achine Learning, Image Registration, Algorithmic Fairness

ticle informations

ps://doi.org/10.59275/j.melba.2025-ae66

ceived: 04/2025, Published 10/2025

rresponding author: dominique.b@northeastern.edu

ecial issue: Special issue on Fairness of AI in Medical Imaging (FAIMI)
est editors: Veronika Cheplygina, Aasa Feragen, Andrew King, Ben Glocker, Enzo Ferrante, Eike Petersen, Esther yol-Antón, Melanie Ganz-Benjaminsen

1. Introduction

Artificial Intelligence (AI) models have achieved expert-level performance in melanoma detection (International Skin Imaging Collaboration, 2020). However, their clinical adoption remains limited due to performance disparities across demographic subgroups, including gender, race, and age (Feng et al., 2023). These disparities are further compounded by differences in melanoma incidence and tumor presentation across the various subgroups (Shao et al., 2022; Marchetti et al., 2021). Addressing these challenges requires comprehensive model auditing strategies that can identify and evaluate subgroup biases before deployment.

One emerging field of AI auditing has been Fairness Benchmarking, characterized by the development of rigorous benchmarking protocols for model assessment under different fairness settings (Han et al., 2023; Weerts et al., 2023; Bellamy et al., 2018). Recent fairness benchmarking methods on clinical data have primarily emphasized *Group Fairness*, which ensures that AI models perform equitably across different demographics. These works utilize a suite of metrics related to the area under the receiver operating characteristic curve (AUROC) to check for balanced predictive ability across subgroups (Zong et al., 2023; Jin et al.,

2024). However, while AUROC captures a model's discriminative power, it does not provide insight into whether a model systematically *over-* or *under-* estimates subgroup risk (Huang et al., 2020). Consider, for instance, a model that predicts the success of vitro fertilization (IVF) – a procedure with variable efficacy designed to assist with conception. Even if such a model accurately classifies successful versus unsuccessful treatments, its clinical utility diminishes if it consistently mispredicts the likelihood of a live birth. Underestimation can increase psychological distress for patients, which in turn impacts specific stages of the IVF process (Zanettoullis et al., 2024). Conversely, overestimation can foster false hope and result in the oversight of critical pre-procedure interventions (Krotz, 2025). To address this limitation, we include an additional metric focused on *calibration*, which quantifies the degree to which a model's average predicted probabilities align with observed event rates (Huang et al., 2020).

Selecting an appropriate calibration metric is critical, as different approaches exhibit different sensitivities to subgroup sizes (Janková et al., 2020; Zhang et al., 2021). Moreover, the choice of relevant subgroups can impose additional biases onto the calibration framework. In order to address these shortcomings, we incorporate a calibration method based on an adaptive-score Cumulative Sum (CUSUM) test as one of our benchmarking tools (Feng et al., 2023). This method enables efficient detection of miscalibration across all subgroups present in the audit dataset, without being restricted to a predefined or limited set of candidate subgroups.

This study evaluates the performance of a state of the art skin cancer detection classification algorithm, winner of the ISIC 2020 Challenge (hereafter ADAE) (Ha et al., 2020), in comparison to the second- and third-place models (hereafter 2nd Place and 3rd Place, respectively) (Pan, 2020; Rota, 2020) when applied to two datasets of varying sizes with a specific focus on subgroups defined by *sex*, *race* (represented in this study by Fitzpatrick Skin Tone; *FST* hereafter), and *age* - specifically, one dataset contains each patient's sex and age, while the other contains all 3. Moreover, we account for intersectionality (i.e., patients spanning multiple subgroups) when evaluating for both AUROC and calibration. The intersectional AUROC analysis is performed by comparing the discrimination of the model between combinations of demographic subgroups, using the DeLong test to assess statistically significant differences in AUROC between and within these combinations. Intersectional calibration analysis is done by applying a score-based CUSUM test and Variable Importance plots (hereafter *VI plots*) to detect and explain miscalibration across combinations of demographic subgroups without requiring predefined subgroup lists. Our experimental results demonstrate that existing models perform comparably to the baseline when applied to datasets with previously unobserved risk factors, a finding that corroborates other benchmarking studies in this domain. Additionally, through our CUSUM calibration test we show that the inclusion or exclusion of patient risk factors results in variations in the model's calibration, such as certain risk factors being well calibrated when included and not calibrated when excluded.

Section 2 reviews related work, Section 3 describes our methodology, Section 4 presents our empirical study and its results, and Section 5 summarizes these results.

2. Related Work

2.1 Fairness in Medical Imaging

A substantial body of literature focuses on defining fairness for algorithmic decision-making systems in healthcare (Chin et al., 2023; Grote and Keeling, 2022; Bærøe et al., 2022). In this context, fairness is often delineated along two principle dimensions: *group fairness* and *individual fairness*.

Individual fairness seeks to guarantee that similar patients – such as ones based on relevant clinical features – receive similar predictions or treatment recommendations (Dwork et al., 2011). It aims to uphold the principle of personalized equity at the cost of a pre-defined similarity metric between individuals, which can be highly nuanced and challenging to obtain (Jui and Rivas, 2024). Given these considerations, we present two frameworks -— group fairness and calibration -— that we believe are most suitable for decision-making systems in healthcare due to their greater flexibility and relaxed assumptions regarding data structure.

Group fairness aims to ensure equitable outcomes across socially salient groups, such as those defined by race, gender, or socioeconomic status, by enforcing independence between a model's predictions and the sensitive attribute (Hutchinson and Mitchell, 2018; Calders et al., 2009; Feldman et al., 2015). Group fairness can be measured using many similarity metrics, such as the classification or misclassification rate of a predictive model amongst different subgroups (Kusner et al., 2018; Hardt et al., 2016); additionally, some definitions of group fairness do not focus on one metric but are designed to allow for a metric of the user's choosing (Lahoti et al., 2020). However, efforts to enforce group fairness can introduce trade-offs with model accuracy, particularly when there are underlying disparities in data representation or label quality. In clinical contexts, such trade-offs carry the risk of violating core bioethical principles such as non-maleficence and beneficence (Beauchamp, 2003; Zafar et al., 2017), by exacerbating disparities in care or introducing harm to less-marginalized populations. Additionally, group fairness metrics are often grounded

in statistical measures, which not only impose inherent limitations but can also lead to mutual incompatibilities among fairness criteria (Kearns et al., 2018; Kleinberg et al., 2016). These factors make the choice of group fairness metric(s) imperative, as it must be done in a way that does not exacerbate the problems that are trying to be mitigated. Motivated by standard model evaluation techniques from clinical contexts (Hong et al., 2023; Alba et al., 2017), we expand our benchmark to not only include group fairness metrics related to AUROC, but calibration as well (**Note:** in line with these techniques, we refer to our AUROC-based analysis as *Model Discrimination* and our calibration-based analysis as *Model Calibration* hereafter).

2.2 Calibration

Calibration is achieved in an AI model when the predicted probability of the model corresponds to the observed event rate. In other words, an AI model is calibrated when, given a population of people, the difference between its estimate and the true outcome is small. Different methods in the literature look to achieve this in different ways while also addressing issues such as variance in subgroup sizes and intersectional group memberships. Specifically, Hébert-Johnson et al. (2017) created an influential method that asks for calibration not only on the entire population, but within specific subgroups as well. Other methods since then look for a similar type of strong guarantee (Kim et al., 2019; Luo et al., 2022). However, rather than checking for calibration amongst every subgroup, this method only checks over a predefined list of subgroups; additionally, these methods each require a high sample complexity in order to achieve any statistical guarantees over these subgroups. Feng et al. define their specific CUSUM test for Model Calibration in a way that allows an auditor to quickly check for miscalibration among all subgroups that appear in the audit dataset and is not limited to a small set of candidate subgroups. This test for strong calibration is also done in one pass of the data. The VI plots that are included as an analysis tool in this method also help to identify when multiple subgroups could be the cause of miscalibration, addressing the issue of intersectional group memberships. These latter two reasons in particular are a significant advantage over methods that only allow for the test of a handful of subgroups at a time. This method also does not require as many samples to guarantee strong calibration as its contemporaries, which typically ask for at least tens or even hundreds of thousands of samples.

2.3 Fairness Benchmarking

There is growing interest in developing standardized benchmarks for assessing AI models in a trustworthy, reproducible, and cross-disciplinary way. Current benchmarks emphasize the most widely-used fairness metrics and algorithms in literature (Han et al., 2023; Weerts et al., 2023; Bellamy et al., 2018). In the medical domain specifically, recent benchmarking efforts have primarily used Model Discrimination to highlight disparities across models. For example, Zong et al. (2023) evaluated a diverse set of models across multiple data modalities (e.g., X-ray, dermatology images, MRI) and explored various model selection strategies using fairness metrics centered around AUROC. Jin et al. (2024) expand upon their work to include the Dice Similarity Score, though its utility is limited to segmentation tasks in computer vision. Our work builds upon these foundations by introducing a complementary perspective: evaluating the role of calibration metrics in fairness assessment. While AUROC-based analysis is a helpful and intuitive way of comparing and understanding model performance, this type of analysis reveals no information about the calibration of these models. Calibration enhances the utility of a model's predictions by ensuring that the predicted probabilities are meaningful and actionable. Without a method for measuring and comparing the calibration of these models, catastrophic decisions could be made in any process that relies on these predictive values.

3. Methodology

3.1 Datasets

The first dataset used in this experiment is the ISIC 2020 Challenge dataset (International Skin Imaging Collaboration, 2020; Kurtansky et al., 2024; Rotemberg et al., 2021). The Society for Imaging Informatics in Medicine and the International Skin Imaging Collaboration (SIIM-ISIC)'s 2020 Melanoma Classification Challenge is hosted on Kaggle and uses a convenience test set of 10,982 public dermoscopy images from six dermatology centers (Barcelona, Spain; New York, United States; Vienna, Austria; Sydney, Australia; Brisbane, Australia; Athens, Greece). The AUC scores for the challenge's final leaderboard were computed over a private, held-out portion of the dataset, while the AUC scores in this study were computed over the whole dataset. The full dataset includes 10,982 participants, with 43% identifying as women and 51.6% under the age of 50.

The second dataset is the PROVE-AI dataset (Marchetti et al., 2023), comprised of 603 images prospectively collected at the Memorial Sloan Kettering Cancer Center (MSKCC), 95 of which are of melanoma. The dataset included 603 Participants (53.7% Women, 49.1% Under the age of 65). Due to the lack of samples from FSTs 4-6, we focus our evaluation on FSTs 1 and 2. However, our analysis is extendable to datasets with sufficient samples

for these groups.

3.2 Selected Methods

In total, 3308 teams participated in the SIIM-ISIC 2020 challenge (Kurtansky et al., 2024). To better understand the performance of the highest ranking approach on the experiments we designed, we also included the next 2 highest ranking approaches in our analysis. Detailed descriptions of the 3 highest ranking approaches are given below. To ensure that each method runs correctly, they were tested on the SIIM-ISIC 2020 data according to their respective publicly available scripts on GitHub[1] [2] [3]. To test the discriminative ability and the calibration of each model when being used on a population different from the one it was trained on, we did not re-train the models on the PROVE-AI dataset and instead simply performed inference with the ISIC weights.

3.2.1 All Data are Ext (ADAE) (1st Place, Ha et al. (2020))

ADAE was the top-performing algorithm in the SIIM-ISIC 2020 challenge, achieving an overall AUROC of 0.9490. It employs an ensemble of eighteen Convolutional Neural Network (CNN) models—each trained across five validation folds, resulting in 90 sets of model weights. Sixteen of these models are based on the EfficientNet architecture, while two use ResNet. Additionally, four of the Sixteen EfficientNet models incorporate clinical metadata, including age, sex, anatomic site, and image size. Training sets from previous years of the challenge were included in the training process to mitigate class imbalances within the ISIC 2020 training set. Multiple image augmentations were also applied to prevent overfitting to the training data. The training involved a 5-fold cross-validation process, with the final model being an ensemble of these cross-validated models.

3.2.2 2nd Place, Pan (2020)

This method utilized an ensemble of fifteen EfficientNet models each trained using five-fold cross-validation. Data augmentation techniques were applied during training to reduce the risk of overfitting. The method was initially trained on the 2019 challenge dataset, followed by training on the combined 2019 and 2020 challenge datasets. However, unlike ADAE, this method did not use any metadata during the training process.

1. ADAE: https://github.com/ISIC-Research/ADAE
2. 2nd Place: https://github.com/i-pan/kaggle-melanoma
3. 3rd Place: https://github.com/Masdevallia/3rd-place-kaggle-siim-isic-melanoma-classification

3.2.3 3rd Place, Rota (2020)

This approach employed an ensemble of eight EfficientNet models, each trained using different combinations of input image resolutions (256, 384, 512, and 768). Training also incorporated data from prior years of the challenge (2017–2019), along with hair augmentation techniques and patient metadata. Unlike the other two methods, no cross-validation was performed on any of the models.

3.2.4 Expected Risk Minimization (ERM) (Vapnik, 1999)

In line with other benchmarking work (Zong et al., 2023), we also include a standard ERM algorithm that serves as a baseline to compare performance. For the ERM baseline, we used a ResNet-18 architecture with cross-entropy loss, trained from scratch using the Adam optimizer. Hyperparameters were selected via Bayesian optimization, with the search space including learning rates in the range $[1 \times 10^{-5}, 1 \times 10^{-3}]$, weight decay values of 1×10^{-4} and 1×10^{-5}, and batch sizes of 256, 512, and 1024. No data augmentation techniques were applied and no metadata was incorporated into the training process. We held out 10% of the ISIC 2020 training data as a validation set, which was used for model selection and early stopping. The final configuration uses a learning rate of 1.475×10^{-4}, selected after being trained for 30 epochs with a batch size of 256. Model Selection was done based on the model that minimized validation loss.

3.3 Metrics and Evaluation

3.3.1 Choice of Calibration Metric

Feng et al. (2023) reframe strong calibration as a Score-Based CUSUM test that indicates if there is *at least one* subgroup that is miscalibrated in a dataset (Feng et al., 2023). Let $\tilde{p} : X \mapsto [0, 1]$ be the risk prediction algorithm and $p_0 : X \mapsto [0, 1]$ be the true event rate over some domain $X \in \mathbb{R}^d$, where d is the number of features. For some pre-specified tolerance level δ, define a poorly calibrated subgroup as $A_\delta = \{x \in X : |\tilde{p}(x) - p_0(x)| > \delta\}$. An AI model is **strongly calibrated** if there is no individual from a dataset that is mis-calibrated.

Suppose the dataset is composed of n independent and identically distributed (i.i.d) observations with samples $x_i \in X$ and binary outcome Y_i for $i = 1, ..., n$. Let $\hat{p} : X \mapsto [0, 1]$ be the AI model being audited, and let $\hat{p}_\delta(x_i) = [\hat{p}(x_i) + \delta]_{[0,1]}$ be its prediction on a sample x_i. Feng et al. (2023)'s method involves partitioning the dataset into two groups of size n_1 and n_2 respectively. n_1 samples are used to train a group of K models to predict the true event rate $p_0(X)$ of each sample (these models are called *residual models* in the original paper). These models are trained on the metadata and predictions from $\hat{p}$ for

each sample. The remaining data is used to calculate a potential test statistic using the product of the differences in the observed and predicted values of the residuals, represented as $(Y_i - \hat{p}_\delta(x_i))$ and $(\hat{p}_0(x_i) - \hat{p}_\delta(x_i))$ respectively. Specifically, the chosen test statistic is given by

$$\hat{T}_{n,>}^{(split)} = \max_{k=1,\ldots,K} \frac{1}{n_2} \sum_{i=n_1+1}^{n} (Y_i - \hat{p}_\delta(x_i))\hat{g}_{\lambda,k}(x)\mathbb{1}\{\hat{g}_{\lambda,k}(x) > 0\} \tag{1}$$

where $\hat{g}_{\lambda,k}(x) = (\hat{p}_0(x) - \hat{p}_\delta(x))$ are the predicted residuals from the kth residual model. A model is said to be *miscalibrated* if the largest of these k test statistics exceeds a threshold.

In addition to this CUSUM test, this method uses Variable Importance plots that are generated by measuring how much the test statistic changes when the features that the residual model was trained on are modified. The most important features are the ones that greatly change the final result of the test when they are modified. Through this two-part process of first using the CUSUM test to check if there is *at least* one miscalibrated group, then using the Variable Importance plots to see which groups they may be, this method is able to circumvent the issues of subgroup imbalance and multi-group membership that are often seen in calibration problems.

In order to improve the estimate of the true event rate, we provide intermediate feature embeddings from each $\hat{p}$ being audited as part of the training data to these residual models. For example, when auditing ADAE on the ISIC 2020 dataset, each sample used to train the residual models contained the original metadata as well as intermediate feature embeddings that were extracted while ADAE was making predictions on the ISIC 2020 test set.

3.3.2 Statistical Analysis

Model discrimination was summarized with AUROC, which is a commonly used metric for clinical binary classification. We evaluate the participating models from two aspects of AUC:

1. **Utility**: AUC gap across different models on the same subgroup (i.e. comparing the AUC of ADAE to that of 2nd Place for women); and
2. **Group Fairness**: AUC gap between different subgroups that were evaluated by the same model (i.e. comparing the AUC of women under 50 to that of men under 50 for ADAE).

For each form of analysis, we use Delong's test (DeLong et al., 1988). DeLong's test is a statistical method used to compare the AUROC curves of two or more models. This test helps determine if there is a significant difference between the AUROC values of the models, which can indicate whether one model performs better than another in terms of classification accuracy and discriminative ability. The test is based on the Mann-Whitney U statistic (Mann and Whitney, 1947), which is equivalent to the empirical AUROC (Bitterlich et al., 2003). It involves calculating the variance of the AUROC and using it to assess the statistical significance of the difference between the AUROC values of the models.

DeLong's test for uncorrelated ROC curves was used for utility analyses, and DeLong's test for correlated ROC curves was used for group fairness analyses. The correlated version of DeLong's test for AUROC is used when the same data is used to generate both AUROC curves. It assumes that the predictions from the two models are correlated because they are based on the same data. On the other hand, the uncorrelated version of DeLong's test is used when different sets of data are used to generate the AUROC curves. It assumes that the predictions from the two models are independent because they are based on different sets of data.

The chosen level of significance was 0.05, and analyses were performed in R (R Core Team, 2021).

4. Results

In this section, we first present our results on each dataset for Model Discrimination, followed by Model Calibration (Sections 4.1 and 4.2, respectively).

4.1 Model Discrimination

4.1.1 ISIC 2020

Table 1 summarizes the performance of the four evaluated models at a 95% sensitivity threshold. While all three ISIC 2020 challenge models outperformed the ERM baseline in terms of AUROC, ADAE exhibited the highest number of false positives across subgroups, suggesting a tendency to over-diagnose.

Tables 2 and 3 present AUROC comparisons across demographic subgroups. Differences in model discrimination between the top three models were generally small and not statistically significant (all p-values > 0.05), with only a few subgroup comparisons showing marginal differences. In contrast, all three models significantly outperformed ERM across every subgroup, reinforcing their overall superiority in discrimination performance.

These results indicate that while the top-performing models are comparable in overall accuracy, their subgroup-specific behavior—particularly ADAE's lower specificity—warrants further scrutiny in clinical contexts.

Table 1: Performance of the 4 models on the ISIC 2020 dataset, stratified by various demographic factors. The top 3 models outperform ERM overall. ADAE has a lower specificity than the other two models which may suggest that it tends to over-diagnose patients at a higher rate.

Characteristic	Total lesions	Total Melanomas	False Positives at 95% Threshold				Sensitivity at 95% Threshold				Specificity 95% Threshold				AUROC			
			ADAE	2nd Place	3rd Place	ERM	ADAE	2nd Place	3rd Place	ERM	ADAE	2nd Place	3rd Place	ERM	ADAE	2nd Place	3rd Place	ERM
Overall	10982	261	3157	2469	2294	5864	95%	95%	95%	95%	71%	77%	79%	45%	0.949	0.949	0.953	0.866
Age																		
Under 50	5306	67	1601	1113	1427	3002	94%	94%	95%	97%	69%	79%	77%	43%	0.942	0.933	0.949	0.862
Over or Equal to 50	5676	194	1556	1356	867	2862	94%	95%	90%	94%	69%	75%	83%	48%	0.951	0.952	0.949	0.867
Sex																		
Female	4727	90	1307	1018	867	2452	93%	92%	91%	94%	72%	78%	81%	47%	0.946	0.940	0.952	0.872
Male	6255	171	1850	1451	1427	3412	96%	96%	97%	95%	70%	76%	77%	44%	0.951	0.954	0.953	0.863

Table 2: Difference in AUROC for each model on the ISIC 2020 dataset using DeLong's correlated test. Significant ($p < 0.05$, blue) and marginally significant ($0.05 \leq p \leq 0.1$, yellow) results are highlighted. The top 3 models all had similar AUROCs on ISIC 2020, and all were much significantly better than ERM.

	ADAE vs 2nd Place		2nd Place vs 3rd Place		ADAE vs 3rd Place	
Subgroup	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**
Everyone	0.001 (-0.007, 0.008)	0.856	-0.004 (-0.011 0.004)	0.333	-0.003 (-0.011, 0.004)	0.421
Men	-0.003 (-0.011, 0.005)	0.493	0.001 (-0.009, 0.011)	0.856	-0.002 (-0.011, 0.007)	0.694
Women	0.006 (-0.008, 0.020)	0.435	-0.012 (-0.024, 0.001)	0.066	-0.001 (-0.019, 0.007)	0.379
Men, Age Over 50	-0.002 (-0.009, 0.006)	0.518	0.011 (-0.001, 0.022)	0.081	0.001 (-0.001, 0.019)	0.092
Women, Age Over 50	0.000 (-0.018, 0.018)	0.980	-0.012 (-0.026, 0.003)	0.115	-0.011 (-0.027, 0.004)	0.159
Men, Age Under 50	-0.001 (-0.026, 0.024)	0.945	-0.023 (-0.052, 0.006)	0.122	-0.024 (-0.053, 0.006)	0.114
Women, Age Under 50	0.018 (-0.002, 0.038)	0.075	-0.009 (-0.026, 0.008)	0.303	0.009 (-0.007, 0.025)	0.257

	ADAE vs ERM		2nd Place vs ERM		3rd Place vs ERM	
Subgroup	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**
Everyone	0.084 (0.065, 0.102)	0.000	0.083 (0.065, 0.101)	0.000	0.087 (0.068, 0.105)	0.000
Men	0.088 (0.066 , 0.111)	0.000	0.091 (0.07, 0.113)	0.000	0.09 (0.068, 0.112)	0.000
Women	0.074 (0.042 , 0.106)	0.000	0.068 (0.037, 0.1)	0.000	0.08 (0.048, 0.112)	0.000
Men, Age Over 50	0.09 (0.065, 0.114)	0.000	0.091 (0.069, 0.114)	0.000	0.081 (0.057, 0.104)	0.000
Women, Age Over 50	0.075 (0.031, 0.118)	0.001	0.075 (0.031, 0.118)	0.001	0.086 (0.045, 0.128)	0.000
Men, Age Under 50	0.086 (0.035, 0.136)	0.001	0.087 (0.033, 0.14)	0.002	0.109 (0.058, 0.16)	0.000
Women, Age Under 50	0.072 (0.03, 0.114)	0.001	0.054 (0.013, 0.095)	0.010	0.063 (0.016, 0.11)	0.009

Table 3: Difference in AUROC for each subgroup on the ISIC 2020 dataset using DeLong's uncorrelated Test. Each comparison produced a difference that was not significant.

Model	Subgroup	Women AUROC	Men AUROC	P-value
ADAE	Age Under 50	0.955	0.931	0.439
2nd Place	Age Under 50	0.937	0.932	0.892
3rd Place	Age Under 50	0.945	0.955	0.665
ERM	Age Under 50	0.882	0.846	0.385
ADAE	Age Over 50	0.942	0.955	0.486
2nd Place	Age Over 50	0.941	0.956	0.465
3rd Place	Age Over 50	0.953	0.946	0.646
ERM	Age Over 50	0.867	0.865	0.955

4.1.2 PROVE-AI

Tables 5, 6, and 4 show the same respective analysis as Tables 1, 2, and 3, but now applied to the PROVE-AI dataset which unlike ISIC 2020 features information about the Fitzpatrick Skin Tone (FST) of each patient.

The top 3 once again outperformed ERM, and on this dataset we see ADAE has the best performance of the top 3 models in terms of False Positives, Specificity and AUROC. At this 95% threshold ADAE had a higher specificity than 2nd Place (40% vs 26%) and 3rd Place (40% vs 10%). 3rd Place was the model that had the most False Positives for every subgroup. This may suggest that of these 4 models, ADAE is best equipped to perform on patients that are of a different distribution than the one it was trained on.

Table 4: Difference in AUROC for each subgroup on the PROVE-AI dataset using DeLong's uncorrelated Test. One marginally significant ($0.05 \leq p \leq 0.1$, yellow) result was produced from this test, with every other comparison producing a difference that was not significant.

Model	Subgroup	FST I AUROC	FST II AUROC	P-value
ADAE	Men	0.89	0.85	0.67
2nd Place	Men	0.93	0.81	0.09
3rd Place	Men	0.62	0.70	0.66
ERM	Men	0.52	0.70	0.27
ADAE	Women	0.82	0.77	0.76
2nd Place	Women	0.77	0.72	0.64
3rd Place	Women	0.58	0.64	0.74
ERM	Women	0.67	0.74	0.58
ADAE	Age Under 65	0.80	0.72	0.58
2nd Place	Age Under 65	0.75	0.65	0.72
3rd Place	Age Under 65	0.64	0.47	0.50
ERM	Age Under 65	0.73	0.48	0.31
ADAE	Age Over 65	0.86	0.83	0.76
2nd Place	Age Over 65	0.83	0.80	0.84
3rd Place	Age Over 65	0.71	0.77	0.64
ERM	Age Over 65	0.72	0.73	0.94

Table 6 depicts the differences in AUROC model discrimination between the risk models for all FST, sex and age combinations considered. Here we see multiple signif-

Table 5: Performance of the 4 models on the PROVE-AI dataset, stratified by various demographic factors. The performance of each model is generally worse than it was on ISIC 2020. In particular, the performance of the 3rd Place model is close to ERM, which is further exemplified in the other tables of this section.

Characteristic	Total lesions	Total Melanomas	False Positives at 95% Threshold				Sensitivity at 95% Threshold				Specificity at 95% Threshold				AUROC			
			ADAE	2nd Place	3rd Place	ERM	ADAE	2nd Place	3rd Place	ERM	ADAE	2nd Place	3rd Place	ERM	ADAE	2nd Place	3rd Place	ERM
Overall	603	95	303	375	457	445	96%	96%	96%	96%	40%	26%	10%	12%	0.850	0.819	0.704	0.735
Age																		
Under 65	307	32	127	177	248	233	94%	91%	94%	97%	54%	36%	10%	15%	0.853	0.828	0.702	0.676
Over or Equal to 65	296	63	176	198	209	212	97%	98%	97%	95%	24%	15%	10%	9%	0.833	0.784	0.684	0.743
Sex																		
Female	324	32	163	170	201	179	94%	97%	97%	94%	44%	30%	12%	9%	0.804	0.770	0.647	0.720
Male	279	63	140	205	256	266	97%	95%	95%	97%	35%	21%	7%	17%	0.865	0.835	0.720	0.735
FST																		
I	46	16	19	27	33	35	100%	100%	89%	89%	49%	27%	11%	5%	0.856	0.874	0.628	0.568
II	333	57	171	214	253	243	94%	92%	96%	98%	39%	24%	10%	14%	0.831	0.790	0.694	0.723
III	202	19	101	123	154	151	97%	100%	100%	94%	41%	28%	10%	12%	0.871	0.835	0.739	0.800
IV	22	3	12	11	17	16	100%	100%	67%	100%	37%	42%	11%	16%	0.947	0.947	0.684	0.807

Table 6: Difference in AUROC for each model on the PROVE-AI dataset using DeLong's correlated Test. Significant ($p < 0.05$, blue) and marginally significant ($0.05 \leq p \leq 0.1$, yellow) results are highlighted. While ADAE has the best overall performance, it and the other top models sometimes do not perform significantly better than ERM.

	ADAE vs 2nd Place		2nd Place vs 3rd Place		ADAE vs 3rd Place	
Subgroup	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**
Everyone	0.031 (0.009, 0.053)	0.006	0.115 (0.055, 0.175)	0.0001	0.146 (0.087, 0.205)	0.00
FST I Men	-0.044 (-0.169, 0.080)	0.485	0.311 (-0.021, 0.643)	0.066	0.267 (-0.061, 0.594)	0.110
FST II Men	0.044 (0.004, 0.084)	0.031	0.111 (-0.003, 0.225)	0.058	0.155 (0.054, 0.256)	0.003
FST I Women	0.045 (-0.169, 0.260)	0.678	0.197 (-0.065, 0.459)	0.140	0.242 (0.070, 0.415)	0.006
FST II Women	0.051 (0.006, 0.095)	0.026	0.077 (-0.069, 0.223)	0.302	0.128 (-0.0234, 0.279)	0.098
FST I, Age Under 65	-0.035 (-0.163, 0.092)	0.589	0.247 (-0.049, 0.543)	0.102	0.212 (-0.081, 0.504)	0.156
FST II, Age Under 65	0.065 (-0.002, 0.132)	0.057	0.057 (-0.093, 0.208)	0.454	0.122 (0.004, 0.241)	0.043
FST I, Age Over 65	0.038 (-0.128, 0.203)	0.657	0.225 (-0.087, 0.537)	0.157	0.263 (-0.035, 0.638)	0.084
FST II, Age Over 65	0.049 (0.006, 0.093)	0.027	0.096 (-0.022, 0.214)	0.110	0.146 (0.031, 0.260)	0.012
	ADAE vs ERM		**2nd Place vs ERM**		**3rd Place vs ERM**	
Subgroup	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**	**Difference in AUROC (95% CI)**	**P-value**
Everyone	0.115 (0.061, 0.169)	0.000	0.084 (0.033, 0.135)	0.001	-0.031 (-0.107, 0.045)	0.425
FST I Men	0.367 (0.131, 0.603)	0.002	0.411 (0.157, 0.666)	0.002	0.1 (-0.298, 0.498)	0.623
FST II Men	0.149 (0.059, 0.24)	0.001	0.105 (0.024, 0.186)	0.011	-0.005 (-0.143, 0.133)	0.938
FST I Women	0.152 (-0.219, 0.522)	0.423	0.106 (-0.112, 0.324)	0.341	-0.091 (-0.481, 0.299)	0.648
FST II Women	0.032 (-0.093, 0.157)	0.614	-0.019 (-0.144, 0.107)	0.771	-0.096 (-0.221, 0.03)	0.135
FST I, Age Under 65	0.2 (-0.112, 0.512)	0.209	0.235 (-0.059, 0.53)	0.117	-0.012 (-0.491, 0.468)	0.962
FST II, Age Under 65	0.233 (0.102, 0.364)	0.000	0.168 (0.043, 0.294)	0.009	0.111 (-0.072, 0.293)	0.234
FST I, Age Over 65	0.425 (0.195, 0.655)	0.000	0.388 (0.181, 0.594)	0.000	0.162 (-0.208, 0.533)	0.390
FST II, Age Over 65	0.047 (-0.051, 0.144)	0.348	-0.003 (-0.094, 0.089)	0.955	-0.099 (-0.221, 0.023)	0.112

icant differences in terms of model performance. Starting with the comparisons of the top 3 models amongst themselves, ADAE has 4 significantly different results from 2nd Place (Everyone, FST 2 Men, FST 2 Women, FST 1 Age Over 65) and 5 from 3rd Place (Everyone, FST 2 Men, FST 1 Women, FST 2 Under 65, FST 2 Over 65). In all but 2 comparisons ADAE was the algorithm that had superior AUROC, highlighting that although ADAE did not perform as well on this dataset as on ISIC 2020, it still handles this new data better than its two closest competitors. Looking at the comparisons of the top 3 models to the baseline shows that all 3 models drop in performance and have multiple subgroups where their discrimination is not significantly better than the baseline; 3rd Place especially has performance that is not significantly better than ERM in any way. This contrasts what was seen when doing the same comparisons on the ISIC 2020 dataset, where each model was significantly better than the baseline amongst every subgroup, and suggests that 3rd Place having the highest AUROC was most likely due to overfitting to the ISIC data.

In terms of AUROC discrimination, there was no significant difference in performance between the different subgroups of FST 1 and 2. In Table 4, when stratified in terms of sex we see AUROCs that range from 58% (3rd Place, Women FST 1) to 93% (2nd Place, Men FST 1) for FST 1, compared to 64% (3rd Place, Women FST 2) and 85% (ADAE, Men FST 2) for FST 2 (All P-values $>$ 0.05). When stratified in terms of age, we see AUROCs for FST 1 that range from 64% (3rd Place, FST 1 under 50) to 86% (ADAE, FST 1 Over 50), compared to 47% (3rd Place, Under 50 FST 2) and 83% (ADAE, Over 50 FST 2) for FST 2 (All P-values $>$ 0.05). These results highlight the poor performance of the 3rd place model, which we explore further in the next section.

4.2 Model Calibration

Table 7 shows the calibration of various versions of the ADAE model using the Score-based CUSUM test of Feng et

al. We tested the calibration of the standard ADAE model (ADAE, Full Ensemble) and two randomly selected two EfficientNet models of the ADAE ensemble, one that uses metadata as a part of the inference process (EfficientNet-M) and one that does not (EfficientNet-NM). The inclusion of the two EfficientNet models was done to determine whether calibration was consistent across individual models in the ADAE ensemble or if it varied depending on architecture and metadata usage.

As part of Feng et al.'s method, an ensemble of Kernel Logistic Regression residual models were trained on the metadata for each respective dataset to predict the true event rate for a given individual in that dataset. The eight residual models evaluated in this experiment varied in their hyperparameter configurations. The first four models used a regularization strength of 1×10^{-3}, with increasing values for the degree of the polynomial kernel approximation: Model 1 used degree 2, Model 2 used degree 3, Model 3 used degree 4, and Model 4 used degree 5. The remaining four models used a regularization strength of 1×10^{-2}, with the same structure: Model 5 used degree 2, Model 6 used degree 3, Model 7 used degree 4, and Model 8 used degree 5. All eight models shared a maximum iteration limit of 2000, used L2 regularization to prevent overfitting, and applied no weighting to zero-labeled samples. Because ADAE was trained previously, we dedicated all data in each respective dataset to training the residual models and testing for strong calibration. Input features to the residual models included demographic information that were available for each dataset (Age and Sex for ISIC; Age, Sex and FST for PROVE-AI) as well as the location of the lesion on the body (Site) and the hospital that the image was collected at (Location). To increase the accuracy of the residual models, we also extracted intermediate features from each model being tested and used those as additional features to train the ensemble of residual models; in the figures below, these intermediate features are labeled as *Feature Embedding (Number)*. We used the *CVScore* variation of this test which performs cross validation by training residual models on subsets of the data and testing for calibration on held-out folds. Our tolerance level for calibration, denoted as δ in the original paper, was set to 0 for all experiments. We provide the technical details of this method in Section 3.

This test computes P-values using a Monte Carlo approach by simulating the distribution of a CUSUM-based test statistic under the null hypothesis of perfect calibration. The P-value is then estimated as the proportion of simulated test statistics that exceed the observed test statistic. For each experimental result discussed below, the P-value was 0.0 so we reject each null hypothesis and assume there's at least one, but possibly multiple, subgroups that are being either Overestimated (i.e. they have a true risk lower than their predicted value) or Underestimated (i.e. they have a true risk higher than their predicted value).

4.2.1 ISIC 2020

Figures 1 and 2 respectively show the tests for overestimation and underestimation in the ADAE model on the ISIC 2020 dataset. The most important features are shown in ascending order on the y-axis of the Variable Importance plots (Figures 1b and 2b). The importance of that feature is defined as the drop in the test statistic after randomly permutating the feature's value, where a larger drop indicates a more important feature.

Focusing first on overestimation, we see the test statistic in Figure 1a quickly rise to a value multiple times higher than that of the one seen in Figure 2a (127.33 vs. 8.45, respectively). This difference in value is due to the difference in the number of samples used in each test; The overestimation case only uses samples that produce positive predicted residuals to build the CUSUM Plot in Figure 1a, while the underestimation case uses samples with negative residuals to do the same. These results suggest that, in the ISIC 2020 dataset, the ADAE model is more prone to overestimating risk than underestimating it.

The case for ADAE overestimating a large number of samples is further supported by the high number of False Positives that we observed in Table 1. For underestimation, we see in Figure 2a that there is an sharp rise in the test statistic across all residual models, followed by a slower increase. These results suggest that there is at least one subgroup identified by the kernel logistic model that is poorly calibrated. The VI plots in Figures 1b and 2b further suggest that the most important features that characterize this poorly calibrated group are age and the original risk prediction. The same features were listed as the most important when running this method on EfficientNet-NM. In the EfficientNet-M model however, age did not show as a strong factor for overestimation, suggesting that the inclusion of patient metadata in the training process may help with controlling over-diagnosis due to age. Interestingly, Both EfficientNet-M and EfficientNet-NM listed various Locations as a high cause for miscalibration while the ADAE model did not list a location in the top 3 features. This difference in the effect of location may be specific to these ResNet models, but not to the other models that make up the ensemble.

4.2.2 PROVE-AI

Figures 3 and 4 respectively show the tests for overestimation and underestimation in the ADAE model on the PROVE-AI dataset. Similar to what we saw with ADAE in our Model Discrimination analysis, the AUROC of each residual model in the ensemble decreased. Looking at

Table 7: The results of Feng et al.'s calibration method on various versions of the ADAE model. Overall, it is shown across both datasets that ADAE is most likely miscalibrated in terms of age more than any other feature, meaning that if a person is to get a risk score too low or too high, their age plays a factor in this mis-assignment. The original risk prediction ('Prediction') is ignored in our analysis since the prediction is inherently tied to the outcome of this calibration experiment.

		Underestimation		Overestimation	
Model	Dataset	Test Statistic	VI Ranking	Test Statistic	VI Ranking
ADAE	ISIC 2020	8.45	Prediction (1), Age (2)	127.33	Prediction (1), Age (2), Sex (3), Location: Athens (5)
	PROVE-AI	9.51	Prediction (1), Age (2), Site: Lateral Torso (3), Sex (6)	9.26	Prediction (1), Age (2), Sex (3), Site: Lower Extremity (4)
EfficientNet-NM	ISIC 2020	14.04	Prediction (1), Age (2), Location: Athens (3), Sex (4), Location: New York (5)	21.15	Prediction (1), Age (2), Sex (4)
	PROVE-AI	10.24	Prediction (1), Age (2), Sex (3)	10.53	Prediction (1), Age (2), Site: Upper Extremity (3)
EfficientNet-M	ISIC 2020	13.63	Prediction (1), Age (2), Location: New York (3), Sex (4)	17.35	Prediction (1), Location: Barcelona (2), Sex (4)
	PROVE-AI	11.18	Prediction (1), Age (2), Sex (3)	11.04	Prediction (1), Age (2), Sex (3), Site: Lower Extremity (5)

both figures, we see the test statistic quickly rise but then smoothen out, suggesting again that there is one subgroup identified by the kernel logistic model that is poorly calibrated. The VI plots in Figures 3b and 4b further suggest that the most important features that characterize this poorly calibrated group are age and the original risk prediction. The same features were listed as the most important when running this method on EfficientNet-NM and EfficientNet-M. All 3 models also on average listed Sex as a higher cause for miscalibration in this dataset than in ISIC 2020. Somewhat surprisingly, we do not see FST as a strong factor of (mis)calibration.

5. Discussion and Conclusion

5.1 Summary of Results

Existing AI models tended to significantly improve discriminative accuracy for melanoma compared to a baseline Expected Risk Minimization (ERM) algorithm when applied to two separate datasets, one from the ISIC 2020 Challenge, where the models were developed, and the PROVE-AI dataset, which featured risk factors that were not present in the former dataset. Prior models generally performed worse on the PROVE-AI dataset, sometimes to the level of ERM, with the best performance exhibited by the ADAE model, which uses an ensemble of 90 models to make predictions.

The model discrimination of ADAE, developed on data from 3 previous ISIC challenges, was significantly better than the other 2 models, developed on data from 2 previous challenges, when applied to the PROVE-AI dataset. Due to differences in the distribution of each dataset, a decrease in performance as observed from the ISIC 2020 test dataset to prospective studies, is expected. Another potential reason for the difference in performance could be the inclusion of patient risk factors in PROVE-AI that was not present in ISIC 2020, such as FST. Another possible explanation might be an inadequate ability of existing models to generalize to unseen data. Focusing on this last issue and exploring the extent to which AI models can improve model performance across sex, FST, and age groups, a baseline AI model, ERM, was trained and validated using data from a previous ISIC Challenge. Overall, ERM had better discriminative performance on ISIC 2020 and experienced a similar drop in performance to the other models when applied to the PROVE-AI dataset. When compared directly, ERM sometimes offered even better performance than 2nd Place and 3rd Place. These findings are consistent with other studies suggesting that there can be limited benefit in using more complex bias mitigation strategies over ERM (Zong et al., 2023).

Poor discrimination was largely driven by the results in the PROVE-AI dataset. Although event rates vary depending on an individual's age, sex and FST, the ability of ADAE

to risk rank individuals was significantly weaker for FST 1 participants than for FST 2 participants for both sexes, and for both age groups. This inaccurate risk assignment for individuals has the potential for ineffective intervention, where higher-risk individuals do not receive beneficial therapy and lower-risk individuals are over-treated. However, it is important to note this discriminative analysis of FST is not complete due to the lack of available FST data for types 5 and 6. Due to this, we refrain from making a strong statement on the discriminative ability of ADAE on the basis of race and instead leave this form of analysis on a larger, more inclusive dataset as future work.

Using the calibration method of Feng et al., we see that ADAE was surprisingly prone to overestimation on ISIC 2020. ADAE also was calibrated well in terms of FST on the PROVE-AI dataset. When examining two of the ResNet-18 models that make up this ensemble, it was shown that different predictive factors led to overestimation and underestimation in the two models. This shows that the inclusion or exclusion of patient metadata in the training of an AI model does indeed play a role in the final calibration of the model, although more analysis needs to be done to determine whether this role is overall beneficial or harmful to the final result. Future work may look to repeat this study with varying combinations of metadata to gain a better understanding of what factors contribute the most to model discrimination.

In terms of FST, more significant differences were observed in terms of model discrimination than calibration. This may be due to limited representation of darker skin tones in the dataset, the relatively lower influence of FST on predicted probabilities compared to features like age and sex, or the calibration method's sensitivity favoring features with stronger statistical signals or larger subgroup sizes. Each form of analysis can be caused by similar issues during the model's creation but can lead to different consequences. Observing multiple differences in model discrimination and calibration implies that the subgroups included in the model training not only require different amounts of information to obtain accurate predictions, but also that the way in which the model learns to predict based on the information provided may favor one subgroup over another. In terms of model discrimination, these two factors could lead to differences in a model's ability to distinguish between positive and negative samples for a subgroup; an example of this would be the difference in AUC observed for FST 1 and 2 for 2nd Place in Table 3, which shows that 2nd Place is significantly better at distinguishing benign cases from malignant for FST 1 than it is for FST 2. In terms of model calibration however, these two factors could lead to unequal preventive measures being applied to a population. As shown by the example in Section 1, overestimation of risk can lead to unnecessary over-treatment; additionally, underestimation of risk can lead to reduced access to care.

Although the best performing models in our study, particularly ADAE, demonstrate strong performance, they also rely on large ensembles of deep neural networks, which require substantial computational resources for training and inference. In contrast, the ERM baseline, which consists of a single ResNet model trained without metadata, requires significantly fewer resources and achieves competitive performance in certain settings, especially on the PROVE-AI dataset. This raises important questions about the tradeoff between computational cost and model performance. In clinical settings, where real-time decision-making, limited hardware availability, and cost constraints are common, models that are computationally efficient may be more practical—even if they offer slightly lower performance. Additionally, simpler models are often easier to validate, interpret, and maintain, which is critical for regulatory approval and clinical trust. Future work should explore this tradeoff more explicitly by reporting metrics such as training time, inference latency, model size, and energy consumption. Such analysis would be valuable for assessing the feasibility of deploying these models in real-world healthcare environments, particularly in low-resource or point-of-care settings.

Despite some of the results showing worse performance, we do not ignore the potential that AI models provide in this domain. Deep learning models have been highly effective in risk prediction because they can extract highly complex, latent features in high-dimensional data sets. As stated in the previous section, future work should look to understand the effect that training a Deep Learning model with different combinations of demographic (race, sex, and age) and clinical (location, lesion type) information has on its performance in our analysis. Other potential solutions to the performance issues outlined in this section could be a technique such as transfer learning (Zhuang et al., 2020), where an AI model trained using data from one distribution of people would be adapted to a second population with a different distribution of demographics and clinical features. Additionally, a technique such as federated learning could be used to train multiple models at once with data from multiple populations (Kaur et al., 2023).

5.2 Limitations

This study was designed to benchmark new and existing techniques for comparing melanoma detection models. In this section, we acknowledge some limitations.

First, we used Fitzpatrick Skin Type (FST) as a proxy for skin tone in evaluating model fairness. However, we acknowledge that FST is not equivalent to race. FST is a dermatological classification based on skin's response to ultraviolet radiation and does not always capture the broader so-

cial, cultural, or systemic dimensions associated with racial identity. Therefore, we caution against interpreting our subgroup analyses as comprehensive assessments of racial fairness. Future work should incorporate more nuanced and demographic data to better understand and address racial disparities in dermatologic AI systems.

Additionally, there was little to no FST data available for 4 of the 6 categories. This prevented us from performing more comparisons of AUC, and therefore limited the experimental results we obtained in terms of FST. Despite this, the analysis we did with FST 1 and 2 did provide an avenue for future work that can be explored further once a more balanced dataset is available.

Along this same line of limitations, only two datasets were used in this experiment. While this is equivalent to the number of skin cancer datasets that were used in similar benchmarks (Zong et al., 2023), this again highlights how future work along this line can redo an experiment of this type with more data and datasets in order to better understand the performance of these top ranking models.

Another limitation lies in the heterogeneity of the evaluated models. The top three ISIC 2020 entries differ not only in ensemble size but also in underlying architecture (e.g., EfficientNet vs. ResNet) and metadata usage. These differences introduce confounding variables that make it difficult to isolate the specific factors contributing to performance or calibration differences. A more controlled comparison—e.g., training a single architecture under varying conditions (with/without metadata, single model vs. ensemble)—would help disentangle these effects. While such an analysis may be beyond the scope of this study, acknowledging this variability is important for interpreting the results.

Furthermore, the models evaluated in this study are large ensembles of CNNs trained across shared cross-validation folds, each with varying parameters. In particular, the ADAE model evaluated in this study is an ensemble of 90 models (Eighteen CNNs each with Five-fold cross validation), with and without metadata. While this ensembling strategy likely contributes to stronger performance, it also introduces complexity that may obscure the behavior of individual models and complicate interpretability. The shared cross-validation setup may limit reproducibility and hinder a more granular understanding of how fairness manifests across the ensemble. We address this by examining two models of this ensemble in our calibration experiments, but we acknowledge that future work could benefit from evaluating simpler or more transparent base models to better isolate the effects of discrimination and calibration techniques.

Finally, the limitations of the calibration method used: the method assumes that residuals can be accurately predicted and that changepoints in residuals correspond to poorly calibrated subgroups. If these assumptions do not hold, the method's reliability may be affected. Future work could look at measuring the uncertainty of the predicted residuals to ensure that the final results are as accurate as possible.

Additionally, while this calibration method (based on the CUSUM test and Variable Importance plots) is designed to detect miscalibration across combinations of demographic subgroups, our results consistently identified only one subgroup as miscalibrated per test. This may reflect limitations in the method's sensitivity to intersectional effects, especially when subgroup sizes are small or when signal strength is diluted across overlapping identities. Additionally, the datasets used in this study may lack sufficient representation of certain intersectional subgroups (e.g., older women with darker skin tones), which could hinder the detection of nuanced calibration disparities. Future work should explore calibration methods with enhanced sensitivity to intersectional miscalibration and apply them to larger, more demographically diverse datasets to better capture these effects.

5.3 Conclusion

While AUROC-based discrimination measures are the more popular way of comparing a model's performance to another, evaluating the calibration of its predictions is also important, especially in clinical applications.

We conducted an extensive benchmarking of existing melanoma detection models to better assess subgroup performance in relation to dataset composition and subgroup size. By evaluating Model Discrimination and Model Calibration metrics, we identified key discrepancies and limitations in the top 3 algorithms for the ISIC 2020 Challenge with respect to fairness, demonstrating the promise of calibration as an auxiliary fairness metric for dermatology algorithms.

Our findings indicate that existing models sometimes perform to the level of a baseline model when applied to datasets with previously unseen risk factors. Additionally, the inclusion or exclusion of patient risk factors causes discrepancies in the final performance of a model, although more information is needed to determine exactly in what way. These results highlight the continued need for more extensive metadata collection from subgroups of interest to achieve dermatologist-level classification performance in this task. Overall, this study serves as a guideline for additional development and auditing of melanoma detection models, as well as model discrimination and calibration methods that can comprehensively evaluate them.

6. Acknowledgments

Funding for this work was provided by NIH/NCI grants U24-CA285296, U24-CA264369, R01CA293974; United States Department of Defense grants HT94252410552, HT94252410553, HT94252410554, HT9425-23-1-0848; NIH/NCI Cancer Center Support grant P30 CA008748; and the Burroughs Welcome Fund Innovation in Regulatory Science Award.

References

Ana Carolina Alba, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, P. J. Devereaux, Thomas McGinn, and Gordon Guyatt. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA*, 318(14):1377, October 2017. ISSN 0098-7484. . URL http://dx.doi.org/10.1001/jama.2017.12126.

T L Beauchamp. Methods and principles in biomedical ethics. *Journal of Medical Ethics*, 29(5):269–274, 2003. ISSN 0306-6800. . URL https://jme.bmj.com/content/29/5/269.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. URL https://arxiv.org/abs/1810.01943.

N. Bitterlich, J. Schneider, and E. Lindner. Roc curves - can differences in aucs be significant? *The International Journal of Biological Markers*, 18(3):227–229, July 2003. ISSN 1724-6008. . URL http://dx.doi.org/10.1177/172460080301800312.

Kristine Bærøe, Torbjørn Gundersen, Edmund Henden, and Kjetil Rommetveit. Can medical algorithms be fair? three ethical quandaries and one dilemma. *BMJ Health &; Care Informatics*, 29(1):e100445, April 2022. ISSN 2632-1009. . URL http://dx.doi.org/10.1136/bmjhci-2021-100445.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009. .

Marshall H. Chin, Nasim Afsar-Manesh, Arlene S. Bierman, Christine Chang, Caleb J. Colón-Rodríguez, Prashila Dullabh, Deborah Guadalupe Duran, Malika Fair, Tina Hernandez-Boussard, Maia Hightower, Anjali Jain, William B. Jordan, Stephen Konya, Roslyn Holliday Moore, Tamra Tyree Moore, Richard Rodriguez, Gauher Shaheen, Lynne Page Snyder, Mithuna Srinivasan, Craig A. Umscheid, and Lucila Ohno-Machado. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Network Open*, 6(12):e2345050, December 2023. ISSN 2574-3805. . URL http://dx.doi.org/10.1001/jamanetworkopen.2023.45050.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44 3:837–45, 1988. URL https://api.semanticscholar.org/CorpusID:21877334.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011. URL https://arxiv.org/abs/1104.3913.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD 2015 - Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268. Association for Computing Machinery, August 2015. . 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2015 ; Conference date: 10-08-2015 Through 13-08-2015.

Jean Feng, Alexej Gossmann, Romain Pirracchio, Nicholas A. Petrick, Gene A Pennello, and Berkman Sahiner. Is this model reliable for everyone? testing for strong calibration. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://api.semanticscholar.org/CorpusID:260315933.

Thomas Grote and Geoff Keeling. On algorithmic fairness in medical practice. *Cambridge Quarterly of Healthcare Ethics*, 31(1):83–94, January 2022. ISSN 1469-2147. . URL http://dx.doi.org/10.1017/S0963180121000839.

Qishen Ha, Bo Liu, and Fuxu Liu. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge, 2020. URL https://arxiv.org/abs/2010.05351.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness

benchmark for in-processing group fairness methods. *ArXiv*, abs/2306.09468, 2023. URL https://api.semanticscholar.org/CorpusID:259187750.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses. *CoRR*, abs/1711.08513, 2017. URL http://arxiv.org/abs/1711.08513.

Chuan Hong, Michael J. Pencina, Daniel M. Wojdyla, Jennifer L. Hall, Suzanne E. Judd, Michael Cary, Matthew M. Engelhard, Samuel Berchuck, Ying Xian, Ralph D'Agostino, George Howard, Brett Kissela, and Ricardo Henao. Predictive accuracy of stroke risk prediction models across black and white race, sex, and age groups. *JAMA*, 329(4):306, January 2023. ISSN 0098-7484. . URL http://dx.doi.org/10.1001/jama.2022.24683.

Yingxiang Huang, Wentao Li, Fima Macheret, Rodney Allanigue Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association : JAMIA*, 27:621 – 633, 2020. URL https://api.semanticscholar.org/CorpusID:211555673.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018. URL https://api.semanticscholar.org/CorpusID:53782832.

International Skin Imaging Collaboration. Siim-isic 2020 challenge dataset, 2020. URL https://challenge2020.isic-archive.com/.

Jana Janková, Rajen D. Shah, Peter Bühlmann, and Richard J. Samworth. Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):773–795, May 2020. ISSN 1467-9868. . URL http://dx.doi.org/10.1111/rssb.12371.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qiongsong Yao, Qi Dou, S. Kevin Zhou, and Xiaoxiao Li. Fairmedfm: Fairness benchmarking for medical imaging foundation models, 2024. URL https://arxiv.org/abs/2407.00983.

Tonni Das Jui and Pablo Rivas. Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, 15(8):3095–3125, January 2024. ISSN 1868-808X. . URL http://dx.doi.org/10.1007/s13042-023-02083-2.

Harmandeep Kaur, Veenu Rani, Munish Kumar, Monika Sachdeva, Ajay Mittal, and Krishan Kumar. Federated learning: a comprehensive review of recent advances and applications. *Multimedia Tools and Applications*, 83(18):54165–54188, November 2023. ISSN 1573-7721. . URL http://dx.doi.org/10.1007/s11042-023-17737-0.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. . URL https://doi.org/10.1145/3306618.3314287.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Information Technology Convergence and Services*, 2016. URL https://api.semanticscholar.org/CorpusID:12845273.

Stephen Krotz. IVF Success Rates and Strategies for Women — inovifertility.com. https://www.inovifertility.com/blog/ivf-success-rates-and-strategies-for-women/, 2025. [Accessed 14-04-2025].

Nicholas R. Kurtansky, Clare A. Primiero, Brigid Betz-Stablein, Marc Combalia, Pascale Guitera, Allan Halpern, Jonathan Kentley, Harald Kittler, Konstantinos Liopyris, Josep Malvehy, Christoph Rinner, Philipp Tschandl, Jochen Weber, Veronica Rotemberg, and H. Peter Soyer. Effect of patient-contextual skin images in human- and artificial intelligence-based diagnosis of melanoma: Results from the 2020 siim-isic

melanoma classification challenge. *Journal of the European Academy of Dermatology and Venereology*, December 2024. ISSN 1468-3083. . URL http://dx.doi.org/10.1111/jdv.20479.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018. URL https://arxiv.org/abs/1703.06856.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning, 2020. URL https://arxiv.org/abs/2006.13114.

Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: Metrics and recalibration, 2022. URL https://arxiv.org/abs/2102.10809.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18 (1):50–60, March 1947. ISSN 0003-4851. . URL http://dx.doi.org/10.1214/aoms/1177730491.

Michael A. Marchetti, Adewole S. Adamson, and Allan C. Halpern. Melanoma and racial health disparities in black individuals—facts, fallacies, and fixes. *JAMA Dermatology*, 157(9):1031, September 2021. ISSN 2168-6068. . URL http://dx.doi.org/10.1001/jamadermatol.2021.2215.

Michael A. Marchetti, Emily A. Cowen, Nicholas R. Kurtansky, Jochen Weber, Megan Dauscher, Jennifer DeFazio, Liang Deng, Stephen W. Dusza, Helen Haliasos, Allan C. Halpern, Sharif Hosein, Zaeem H. Nazir, Ashfaq A. Marghoob, Elizabeth A. Quigley, Trina Salvador, and Veronica M. Rotemberg. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (prove-ai study). *npj Digital Medicine*, 6(1), July 2023. ISSN 2398-6352. . URL http://dx.doi.org/10.1038/s41746-023-00872-1.

Ian Pan. kaggle-melanoma/documentation.pdf at master · i-pan/kaggle-melanoma — github.com. https://github.com/i-pan/kaggle-melanoma/blob/master/documentation.pdf, 2020. [Accessed 12-03-2025].

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Cristina Rota. 3rd place solution overview. https://www.kaggle.com/c/siim-isic-melanoma-classification/discussion/175633, 2020. [Accessed 12-03-2025].

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8 (1), January 2021. ISSN 2052-4463. . URL http://dx.doi.org/10.1038/s41597-021-00815-z.

Kimberly Shao, Jette Hooper, and Hao Feng. Racial and ethnic health disparities in dermatology in the united states. part 2: Disease-specific epidemiology, characteristics, management, and outcomes. *Journal of the American Academy of Dermatology*, 87(4):733–744, October 2022. ISSN 0190-9622. . URL http://dx.doi.org/10.1016/j.jaad.2021.12.062.

V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. ISSN 1045-9227. . URL http://dx.doi.org/10.1109/72.788640.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023. URL https://arxiv.org/abs/2303.16626.

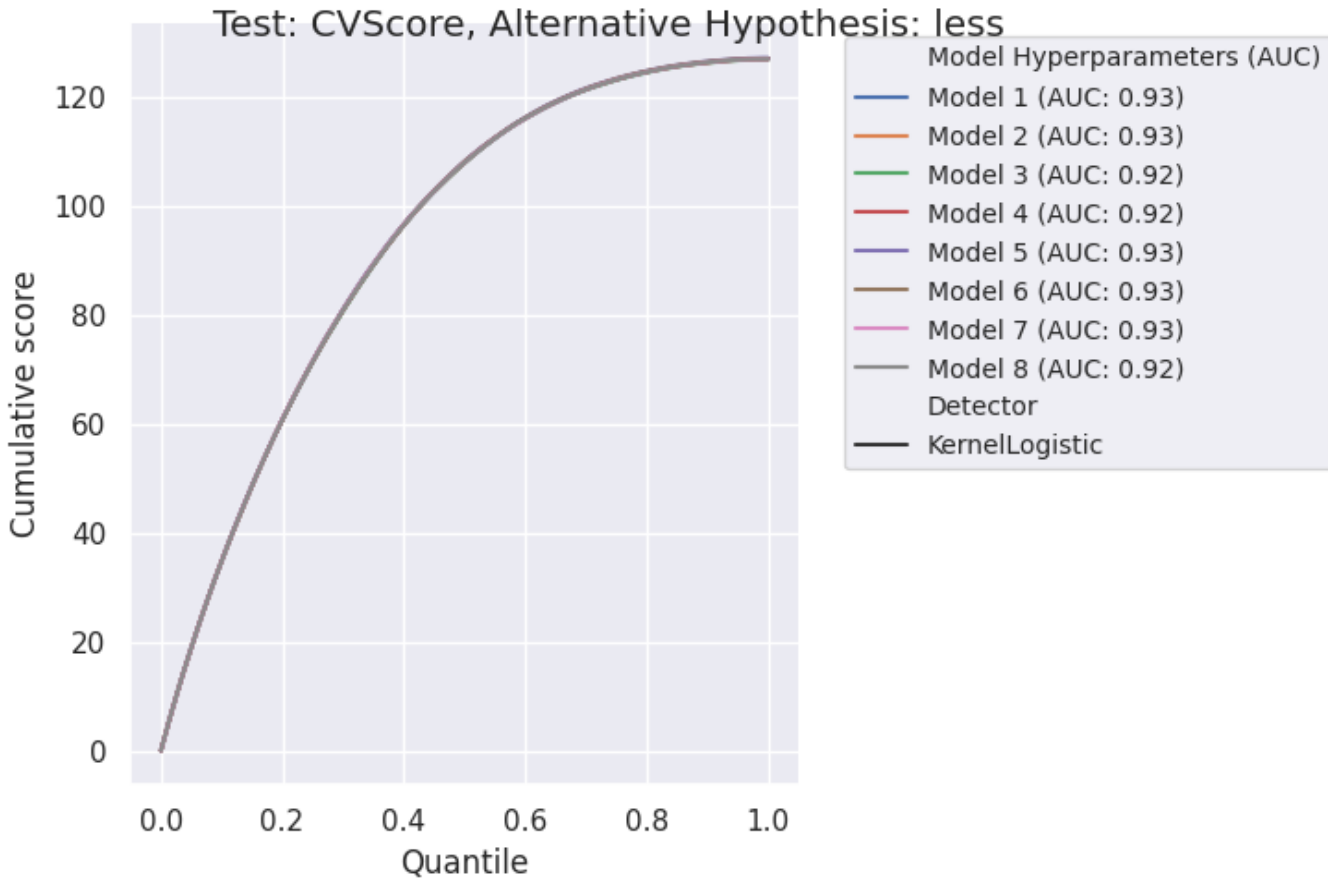
Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/82161242827b703e6acf9c726942a1e4-Paper.pdf.

Anastasia Tsambika Zanettoullis, George Mastorakos, Panagiotis Vakas, Nikolaos Vlahos, and Georgios Valsamakis. Effect of stress on each of the stages of the ivf procedure: A systematic review. *International Journal of Molecular Sciences*, 25(2):726, January 2024. ISSN 1422-0067. . URL http://dx.doi.org/10.3390/ijms25020726.

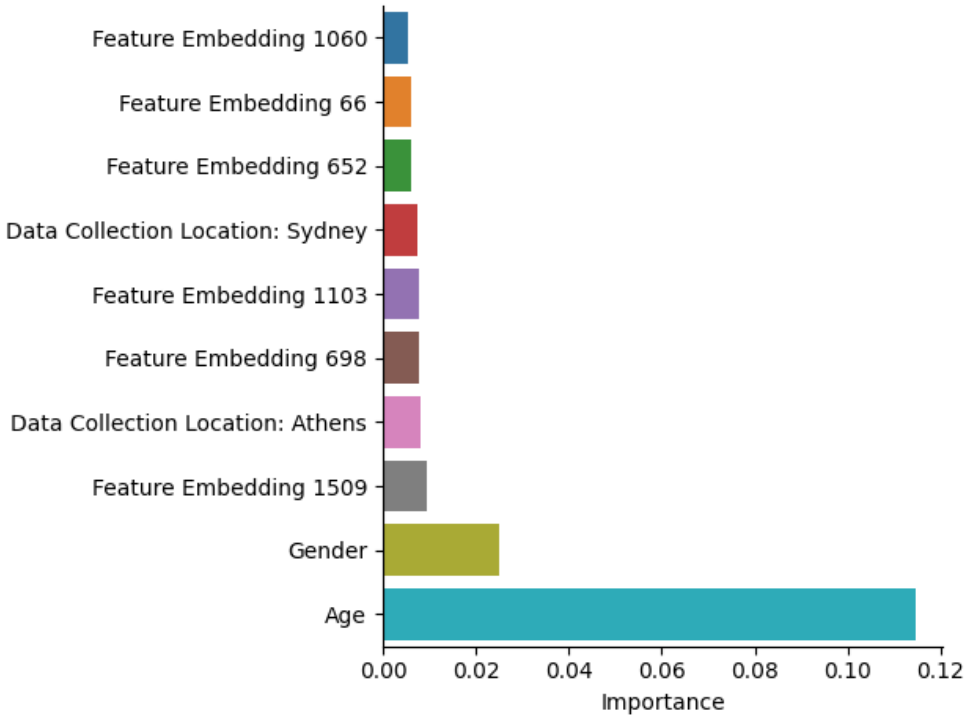
Jiawei Zhang, Jie Ding, and Yuhong Yang. Is a classification procedure good enough?—a goodness-of-fit assessment tool for classification learning. *Journal of the American Statistical Association*, 118(542):1115–1125, November 2021. ISSN 1537-274X. . URL http://dx.doi.org/10.1080/01621459.2021.1979010.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2020. URL https://arxiv.org/abs/1911.02685.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging, 2023. URL https://arxiv.org/abs/2210.01725.

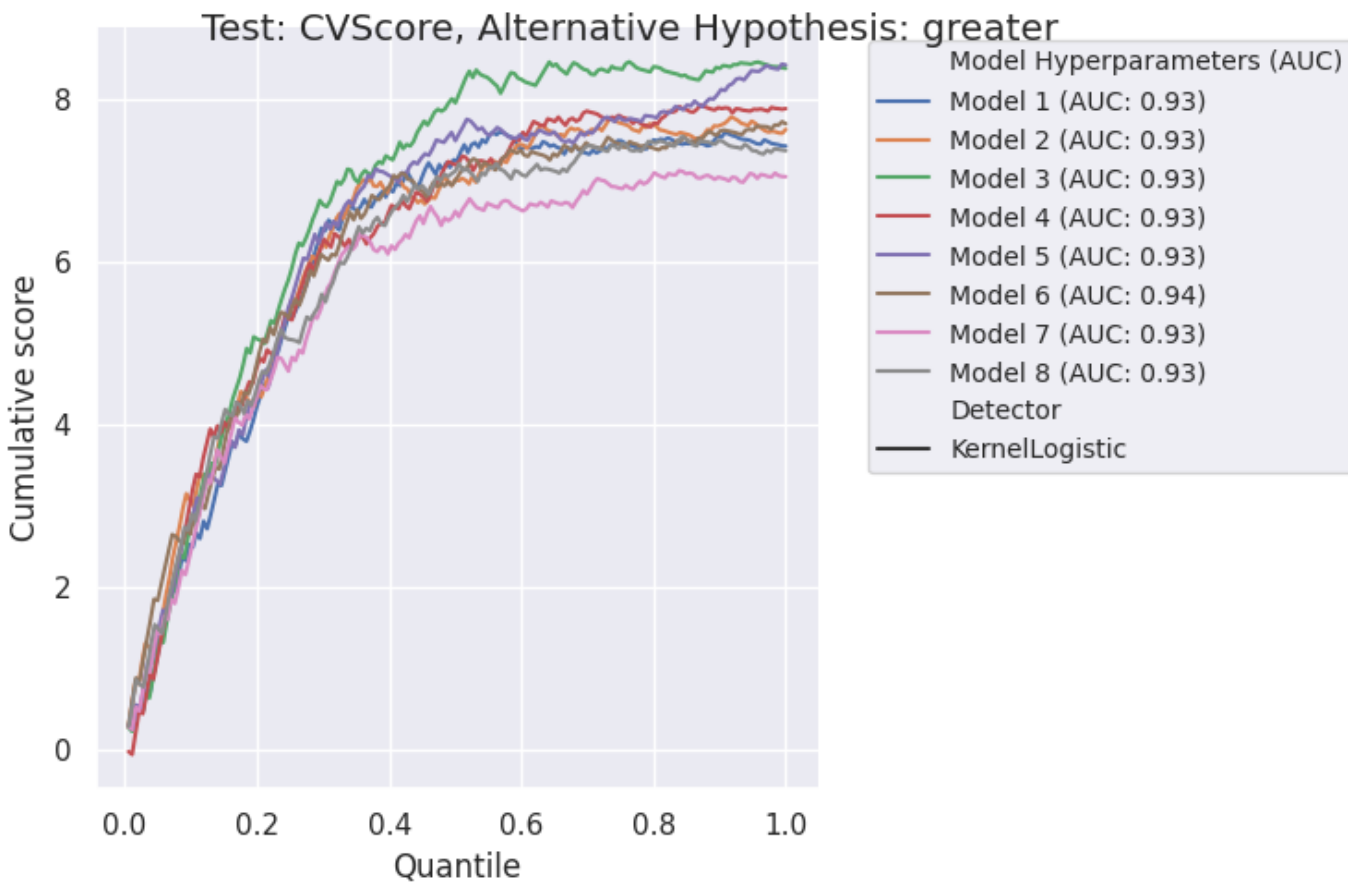


(a) A control chart plotting cumulative score. Here, the ADAE model is being checked for overestimation on the ISIC 2020 dataset.

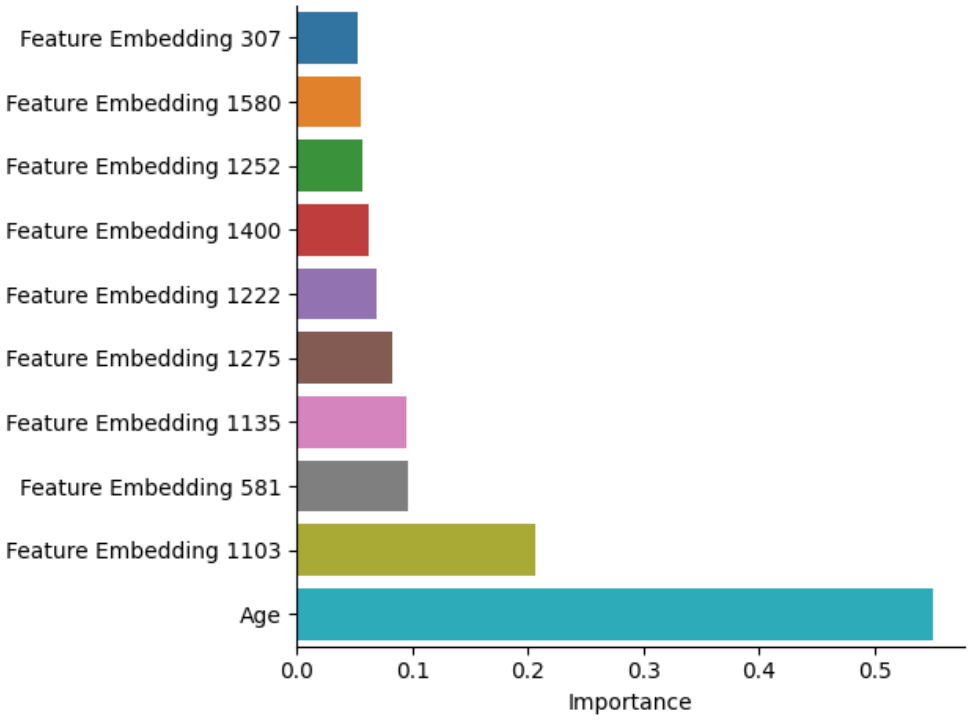


(b) Variable Importance Plot for the top 10 features.

Figure 1: A control chart and Variable Importance plot to check for subgroups who have their true risk overestimated in the ISIC 2020 dataset.

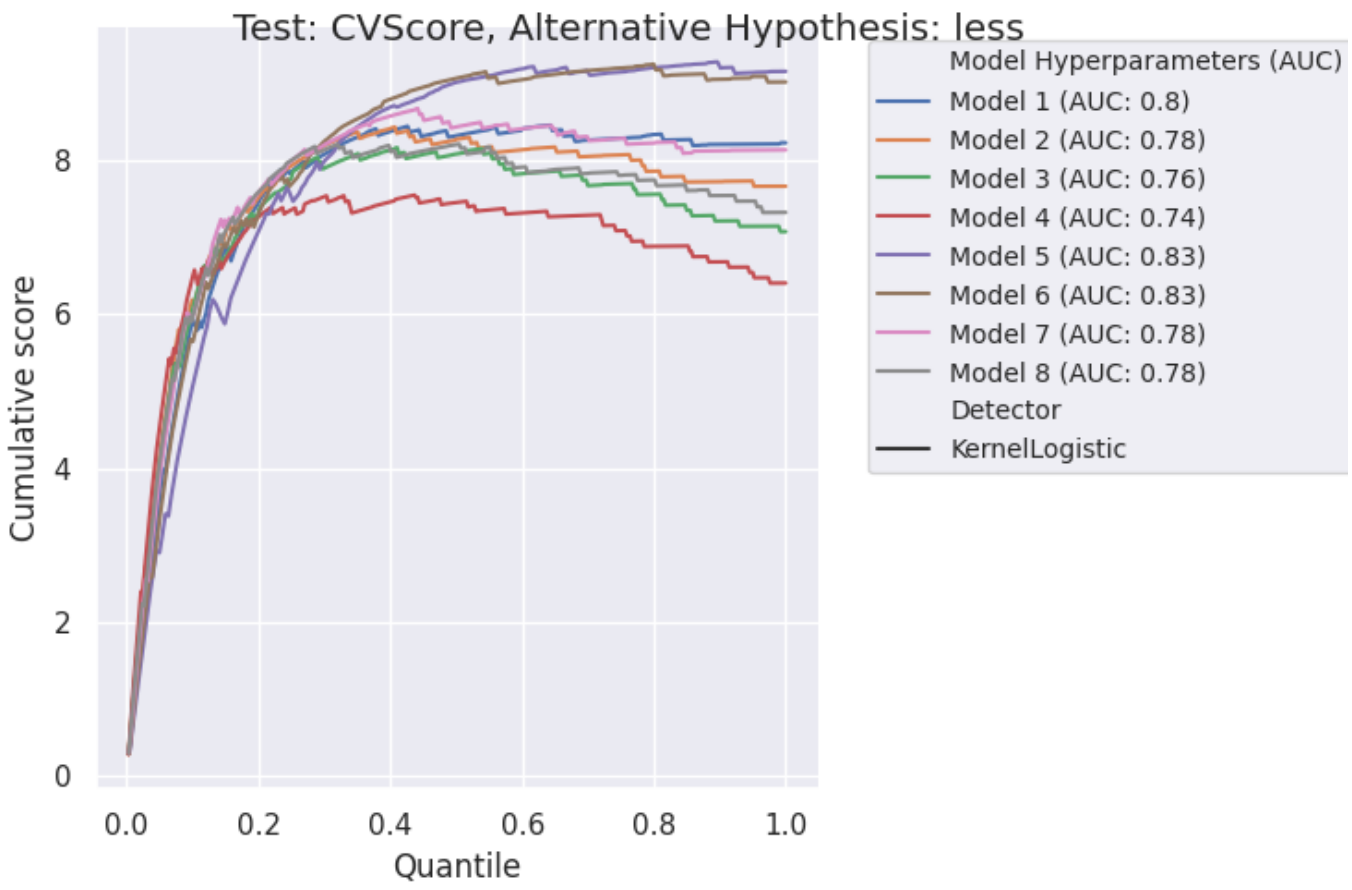


(a) A control chart plotting cumulative score. Here, the ADAE model is being checked for underestimation on the ISIC 2020 dataset.

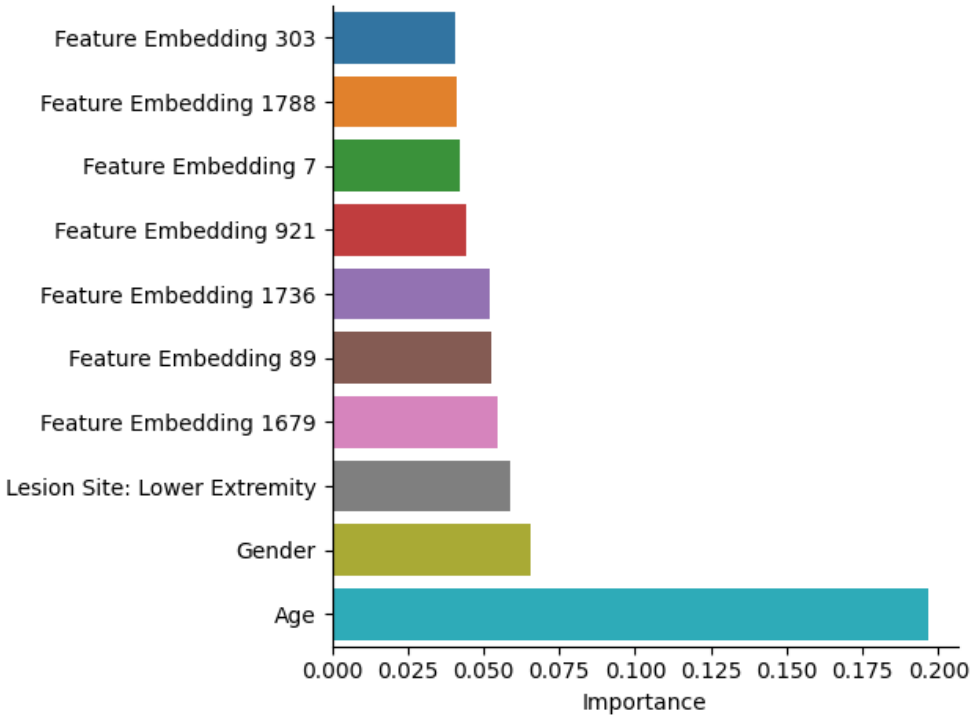


(b) Variable Importance Plot for the top 10 features.

Figure 2: A Control Chart and Variable Importance plot to check for subgroups who have their true risk underestimated in the ISIC 2020 dataset.

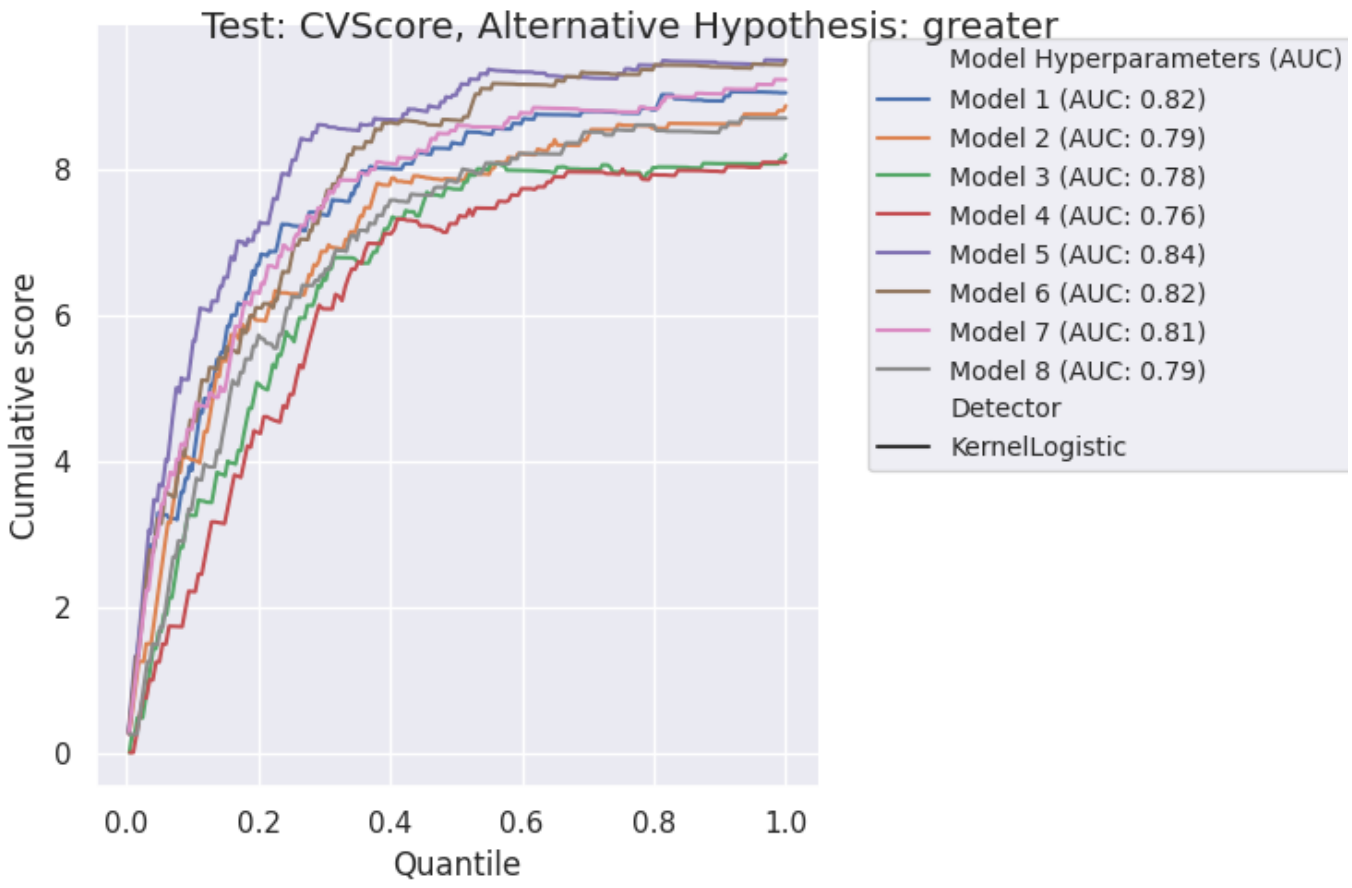


(a) A control chart plotting cumulative score. Here, the ADAE model is being checked for overestimation on the PROVE-AI dataset.

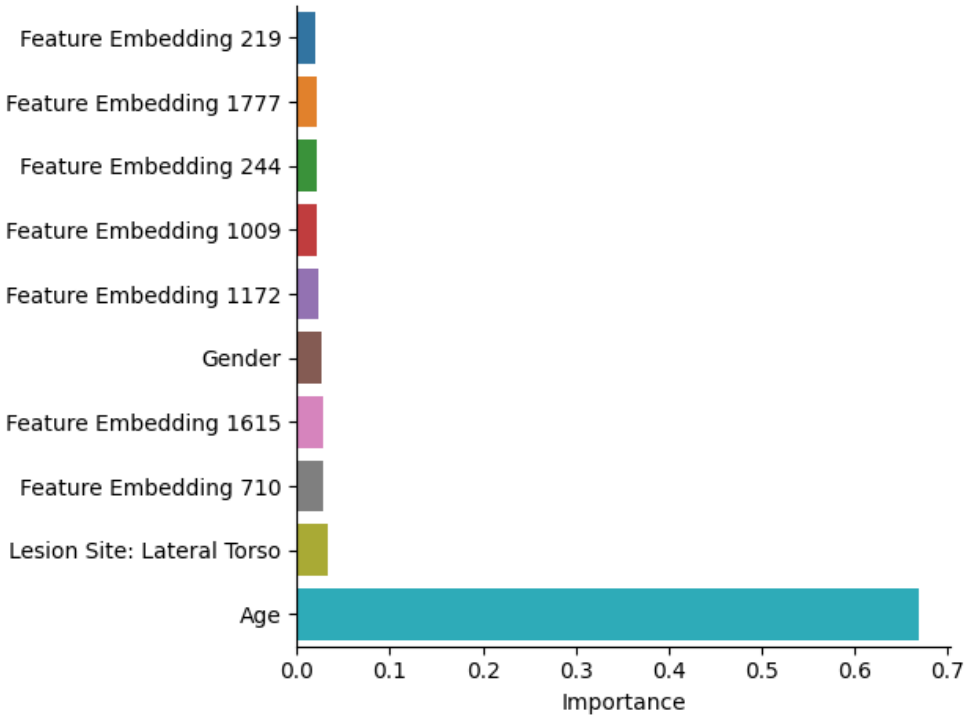


(b) Variable Importance Plot for the top 10 features.

Figure 3: A control chart and variable importance plot to check for subgroups who have their true risk overestimated in the PROVE-AI dataset.



(a) A control chart plotting cumulative score. Here, the ADAE model is being checked for underestimation on the PROVE-AI dataset.



(b) Variable Importance Plot for the top 10 features.

Figure 4: A control chart and variable importance plot to check for subgroups who have their true risk underestimated in the PROVE-AI dataset.