Beyond the Surface: Advancing Skin Cancer Detection Across All Tones

ChiChun Chen
Department of Statistics
University of California, Davis
Davis, CA 95616
ciqchen@ucdavis.edu

Savali Sandip Deshmukh
Department of Computer Science
University of California, Davis
Davis, CA 95616
sdeshmukh@ucdavis.edu

Shivani Sanjay Suryawanshi
Department of Computer Science
University of California, Davis
Davis, CA 95616
ssuryawanshih@ucdavis.edu

Sriharshini D
Department of Computer Science
University of California, Davis
Davis, CA 95616
sdusi@ucdavis.edu

Yehong Qiu
Department of Statistics
University of California, Davis
Davis, CA 95616
yehqiu@ucdavis.edu

Abstract

This project addresses disparities in skin cancer detection caused by biases in machine learning models trained predominantly on lighter skin tones. Using the HAM10000 dataset, we developed models, including FixCaps, Inception V3 and Logit Regression and evaluated them with conformal prediction techniques, to improve classification accuracy and fairness across diverse skin tones. Our findings highlight FixCaps as the most effective model, offering reliable and inclusive diagnostic capabilities. By integrating fairness-aware metrics and robust architectures, this work advances equitable AI solutions for medical imaging. For more details and access to the code, please visit our GitHub repository at https://github.com/Yehongqiu-lab/STA-221-Project.

1 Introduction

In today's world, where technology and healthcare intersect, the need for inclusive and adaptable medical AI tools is crucial. Image classification plays a pivotal role in the medical field, enabling automated analysis of medical images that can assist in diagnostics, improve efficiency, and potentially save lives. Image classification in medical contexts allows healthcare providers to detect abnormalities and make informed decisions faster than traditional methods. One significant area where image classification can make a profound impact is in the detection of skin cancer, a condition that requires early identification to improve patient outcomes.

While automated classification systems have shown promise in skin cancer detection, most models are trained on limited, homogeneous datasets that predominantly feature lighter skin tones. This limitation results in diagnostic disparities, as models often fail to accurately identify skin cancer on individuals with darker skin tones. Expanding classification models to be effective across a range of skin tones is not only scientifically necessary but also ethically crucial. A model trained inclusively

on diverse skin tones could help address healthcare disparities, providing more equitable access to early skin cancer detection for all individuals.

Our project aims to narrow this gap by evaluating image classification model's fairness across skin tones with Conformal Prediction (CP), a distribution-free uncertainty quantification technique on a model's confidence of delivering correct outcomes. By training and testing our models on a carefully curated, labeled dataset representing a spectrum of skin tones, we aim to create a robust, inclusive statistical modeling workflow for skin cancer detection.

Our goals for this research are to:

- Highlight the need for skin tone inclusivity in skin cancer diagnosis to build a fairer, stronger healthcare AI system
- Develop and evaluate several classification models on a skin lesion image set
- Compare different ML models to determine the most effective approach for skin cancer classification in diverse skin tones

1.1 Research Questions

To address the challenges in skin cancer classification across diverse skin tones, we have formulated the following research questions:

1.1.1 RQ1. Classification Effectiveness Across Skin Tones

Which machine learning model achieves the highest accuracy in classifying skin cancer cells across varied skin tones (e.g., lighter vs. darker tones) within a supervised learning framework?

Importance: This question addresses the primary objective of developing a robust model for skin cancer classification that works effectively across skin tones. A model that performs well on diverse skin tones is essential to ensure fair and inclusive diagnostic support for all individuals.

1.1.2 RQ2. Model Comparison in Supervised Learning

How do CNN-based models such as Inception V3 and FixCaps perform compared to other image classification models in supervised learning for multi-class skin cancer classification?

Importance: This question allows us to understand the strengths and limitations of different machine learning models within a supervised learning context, helping identify the most effective architecture for skin cancer detection across diverse skin tones.

1.1.3 RQ3. Fairness and Accuracy in Skin Cancer Detection

What metrics best capture the fairness and accuracy of skin cancer detection models across diverse skin tones? To achieve this, we might use multiple measures, including the aforementioned CP to evaluate a model's reliability.

Importance: This question examines the broader impact of our model on fairness in healthcare, specifically by ensuring equitable diagnostic performance across all skin tones. By identifying fairness metrics, we aim to contribute to the development of ethically sound AI solutions in healthcare.

2 Dataset

The HAM10000 dataset(2) is widely recognized and extensively used in medical image analysis research. It offers a large and diverse set of dermatoscopic images, which are crucial for developing robust machine learning models. The inclusion of various skin tones within the dataset helps address biases and ensures the model's applicability across different demographics.

This dataset comprises 10,015 images of skin lesions. Key columns include 'imageid' (unique identifier for each image), 'dx' (diagnostic label, such as melanoma, nevus, or seborrheic keratosis), 'dxtype' (type of diagnostic procedure, like histopathology), 'age' (age of the patient), 'sex' (gender of the patient), 'localization' (anatomical site of the lesion), and 'path' (path to the image file).

The dataset includes seven categories of skin lesions: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions, and dermatofibroma (See figure 1). This categorization is vital for our primary prediction goal, which is to classify skin lesions accurately.

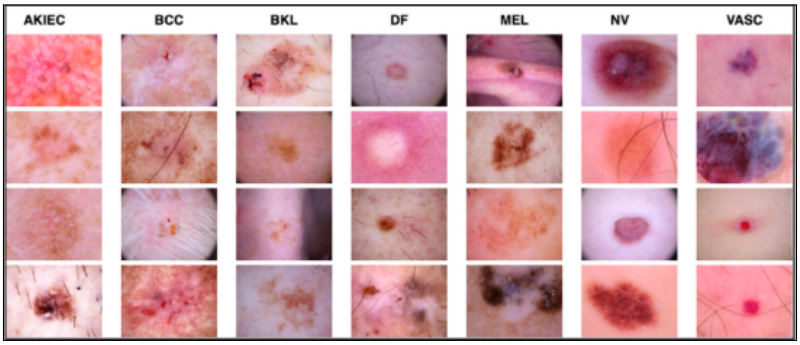

Figure 1: *HAM10000 Dataset*

We sourced the HAM10000 dataset from publicly available platform Kaggle(10). Our utilization strategy involves leveraging the image data of skin lesions as predictor variables and the classification labels (e.g., the seven categories of skin lesions) as the target variable. The dataset's comprehensive nature and its inclusion of diverse skin tones make it an ideal choice for this project.

This is how the original dataset distribution (See figure 2) across classes looks like:

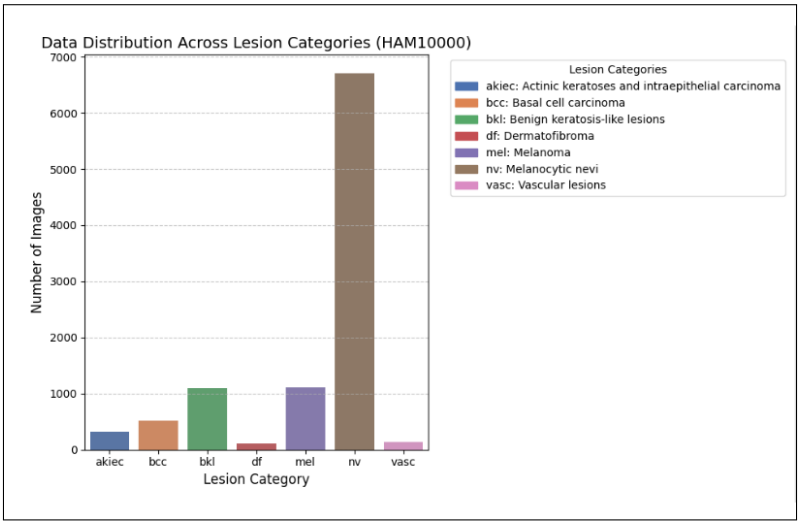

Figure 2: *Data distribution*

It can be inferred that the classes are imbalanced. nv (Melanocytic nevi) dominates with 6705 images (67%) and classes like df (Dermatofibroma) and vasc (Vascular lesions) have only 115 (1%) and 142 (1.4%) images, respectively. This can lead the models to perform poorly on underrepresented classes, such as df and can also be biased towards the dominating classes like nv. Therefore, we perform data augmentation to ensure fair and effective model training.

2.1 Preprocessing and Augmentation:

Preprocessing the images involved resizing images to a consistent dimension, normalizing pixel values to enhance model performance, and augmenting the dataset (See figure 3) (using rotation, flipping, zooming and cropping) to create a balanced dataset of 56,000 images across seven lesion categories (See figure 4) with tensorflow keras ImageDataGenerator library. Labels were numerically

encoded, and metadata is cleaned to handle missing values, drop irrelevant columns like patient details and to ensure consistency.

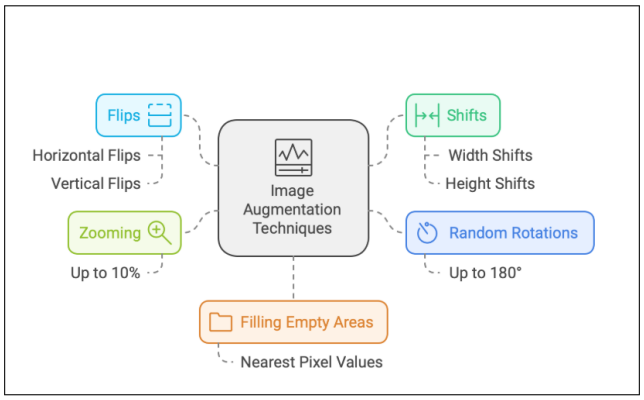

Figure 3: *Data augmentation*

To assess the model's generalization performance, we used the benchmark ISIC (International Skin Imaging Collaboration) dataset(1) as the test set, which has about 240 images. Due to its high-quality annotations and varied coverage of skin lesion types, this dataset is well-known in the field of dermatological image analysis. The ISIC dataset(10) provides an effective testing ground for evaluating the model's capacity to generalize outside of the training data because it contains images from a variety of sources and patient demographics. The data set also includes detailed metadata, such as lesion type, demographics of the patient, and lesion location.

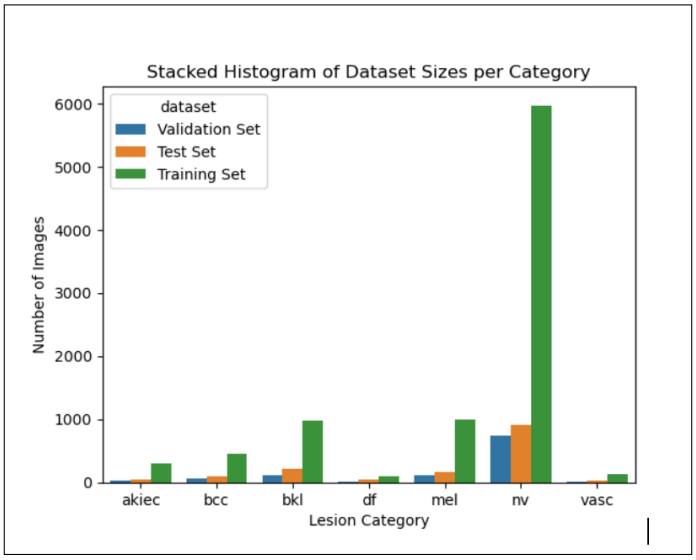

Figure 4: *Data Distribution before augmentation*

This is how the data distribution looks after augmentation:

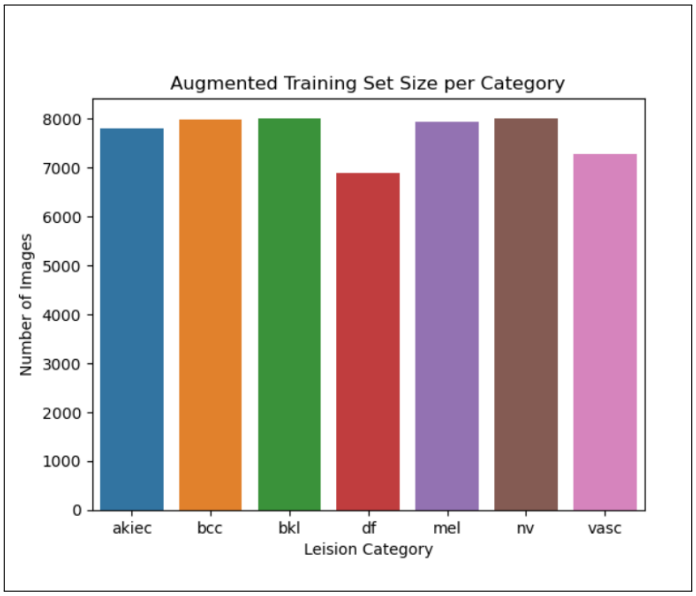

Figure 5: *Augmented Data Distribution*

3 Literature Review

3.1 FixCaps: An Improved Capsules Network for Diagnosis of Skin Cancer(3)

We chose the following paper because FixCaps[1] is currently the world-leading DL model on HAM 10K. The paper presents FixCaps, a modified capsule network designed for classifying skin lesions, trained on the HAM10000 dataset. FixCaps enhances CapsNet by introducing a larger kernel size in the initial convolutional layer (See figure 6), a convolutional block attention module (CBAM) between the convolutional layer and the capsule layer, reducing the losses of spatial information caused by convolution and pooling, and a group convolution used to avoid model underfitting in the capsule layer. These add-ons improve the network's ability to capture spatial hierarchies, which are essential for detailed skin lesion images. Implemented with data augmentation techniques, the FixCaps model outperformed prior architectures in classification accuracy, achieving up to 96.49% on the HAM10000 dataset(2). It also reduces computational costs compared to previous methods with fewer parameters, making it efficient for clinical applications.

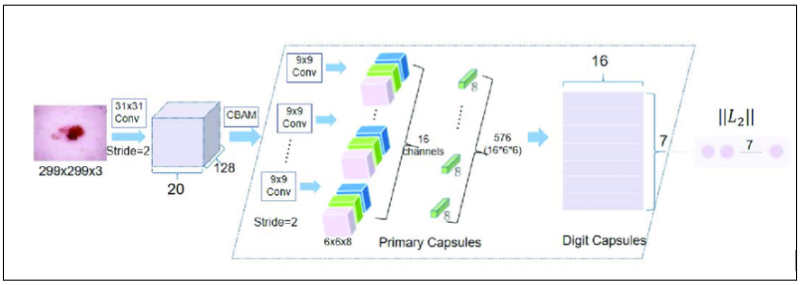

Figure 6: *FixCaps Architecture*

In developing their FixCaps model, the authors assume several key points:

- Distribution Similarity: They assume that the data for training, validation, and testing is under the same population distribution.

- Spatial Hierarchies in Skin Lesions: They assume that skin lesion images contain spatial hierarchies, such as texture and patterns, which capsule networks can better capture than traditional CNNs which use pooling to reduce spacial dimensions while discarding some positional information.

These assumptions guide their architectural choices, focusing on spatial information and computational efficiency.

3.2 Fair Conformal Predictors for Applications in Medical Imaging(4)

We chose the following paper[3] because we want to evaluate the reliability of machine learning models on sub-group accuracy, and Conformal Prediction (CP) (See figure 7) provides distribution-free coverage guarantees regardless of the input data and the implemented machine learning model. Although CP can only give us the average accuracy/coverage guarantees on the aggregate data regardless of subgroups, it can be easily extended to provide equalized coverage guarantees within each subgroup as long as we can use a group attribute to distinguish one from the other. And this extension is called Group-Balanced Conformal Prediction (GBCP). Since our purpose is to find out a fairer ML model that can provide convincing skin cancer classification on any skin tones, GBCP is an important tool to use for evaluating our model's reliability. The paper extended CP to Group-Balanced Conformal Prediction (GBCP) to produce prediction sets that have equal error rates across certain subsets of the data, i.e., across different patient groups (different skin tones).

The authors adapted adaptive prediction sets (APS) into Group-Balanced Conformal Prediction (GBCP) to achieve equalized coverage for a pre-specified marginal coverage level, α, on the skin-tone attribute group A. They then constructed a ResNet-18, pretrained on ImageNet, as their deep learning image classifier and experimented primarily with a dermatology photography dataset used for skin lesion classification, called Fitzpatrick 17k, which includes 7 skin tone types and 114 conditions (with only 11 being malignant) to test their model's fairness across demographic groups.

The empirical study yielded promising results: the GBCP algorithm provided a fairer way of looking into ML models while giving out evaluation on models' reliability. Specifically, the cardinality of the prediction sets produced by GBCP is a direct metric on model's confidence: the larger it is, the less confident the model is; and vice versa.

The authors' key assumptions include:

- Sufficient Data: They assume that the dataset is large enough that each subgroup can provide sufficient data for separate calibration, which is crucial for GBCP.
- No Distribution Shift: Conformal prediction sets are not reliable if there is a distributional shift of the input variable or the label distribution between the calibration data and test data. The authors assume that there is no distributional-shift.
- Sufficient Number of Classes: The authors assume that the number of classes/labels is large enough to provide useful information within a prediction set. For instance, if a medical image classification is only binary, there are only three possible prediction sets, which are both classes individually and the combination of both.

This problem remains when the number of classes is few, so the practical value of conformal predictions in settings with few classes or categories can be low due to the coarse resolution of prediction sets.

In high-stakes applications like medical imaging, conformal prediction (CP) is a technique that offers accurate uncertainty quantification and calibrated prediction ranges. A scoring function $S(x_i)$, which is obtained from the outputs of a base model on a calibration dataset, is computed at the start of the process. This score indicates how likely or confident a prediction is and is calculated by comparing it to the samples in the calibration data. There are many choices on calculating the scores, but generally, the larger the score, the greater the disagreement between the ground truth and the prediction. The empirical quantile ($\hat{q}$), which acts as a threshold for choosing which predictions to include in the prediction set of the validation data, is computed using the scores from the calibration data. For instance, the threshold at which prediction sets meet the required level of confidence is defined by the $(1-\alpha)$-quantile (for instance, the 95th percentile for $\alpha = 0.05$).

Consider, for a test skin lesion image, the conformal prediction method generates a prediction set, $C = \{\text{'bcc'}, \text{'mel'}\}$, which includes the possible classifications for the lesion with a specified confidence level. The true label for this sample is 'bcc' and it is included within the prediction set, ensuring the model's prediction meets the reliability guarantees of conformal prediction. Conformal prediction provides calibrated uncertainty quantification by providing a set of potential outcomes rather than a single label, improving the reliability and transparency of medical diagnostics.

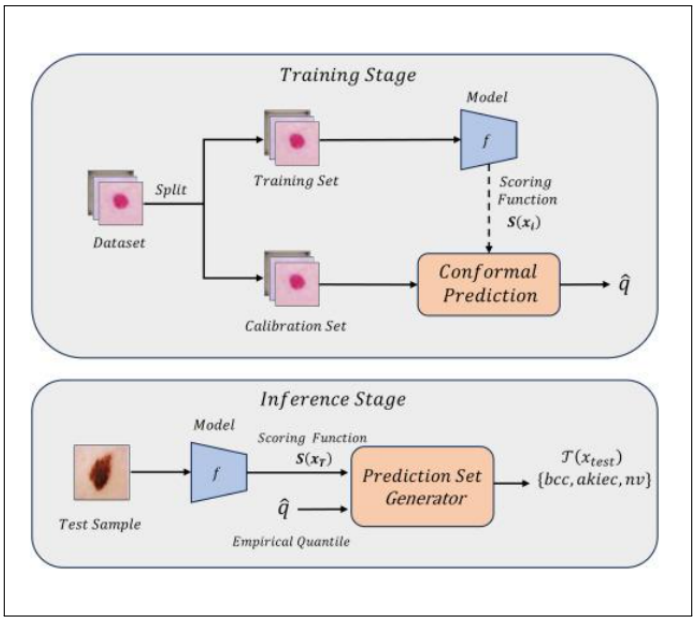

Figure 7: *Conformal Prediction Architecture*

In the training stage, a nonconformity or conformity score is determined for every calibration sample by calculating the distance between the true label and the projected confidence of the model. The threshold for creating prediction sets with assured error rates is the $(1 - \alpha)$-quantile of the calculated scores.

In the inference stage, the trained model uses the scoring function $S(x_{\text{test}})$ to produce a confidence score for the test image.

The empirical quantile, $\hat{q}$, is compared to the score. The prediction set $T(x_{\text{test}})$ will then contain every possible class that satisfies the marginal coverage.

3.3 Conformal Prediction(6)

We chose to implement CP instead of Group-Based Conformal Prediction (GBCP), due to some challenges we came across when trying to implement GBCP.

When the calibration set is divided into subgroups for group-based adjustments, the size of each subset is decreased, which may lead to less reliable quantile estimation and lower coverage guarantees for smaller demographic groups. The need to balance coverage across all subgroups may lead to a trade-off, where overall marginal coverage across the population is reduced compared to standard CP. Also, GBCP is resource-intensive, particularly for large-scale or high-dimensional datasets, as different nonconformity scores and thresholds must be calculated for each subgroup.

As mentioned above, there are advantages to using CP like its model-agnostic nature, uncertainty quantification, and the ability to set confidence thresholds (e.g., 95%) and this ensures effectiveness in handling imbalanced data and fair error rates across classes.

The calculation of the conformal score $S(x_i)$ is often based on the softmax scores of underlying models. A specific example is to let it be $1 - \text{Softmax}_{Y_{\text{truth}}}$, where $\text{Softmax}_{Y_{\text{truth}}}$ is the softmax probability assigned to the true label of the calibration point and

$$\hat{q} = \text{Quantile}_{1-\alpha}\left(\{S(x_i)\}_{i=1}^{n}\right)$$

Xtest is the test sample, and $C(X_{\text{test}})$ is the conformal prediction set, which includes all classes y that meet the condition: $S(x_{\text{test}}, y) \leq \hat{q}$.

True label will be included in the *prediction set* with a probability of at least $1-\alpha$, achieving *marginal coverage*.

$$P\left(Y_{\text{test}} \in C(X_{\text{test}})\right) \geq 1 - \alpha.$$

4 Methodology

4.1 Workflow

The methodology for this project consists of a structured sequence of steps, designed to ensure development and evaluation of multiple robust machine and deep learning models. We reviewed existing literature and research to understand current best practices and gain insights into the dataset's structure, patterns, and features. First, the HAM10K dataset was split into training and validation to facilitate effective model training and performance evaluation. Next, data augmentation techniques were applied to balance and enhance the training set, addressing potential biases and improving diversity. Machine learning models were trained using the prepared training data, optimizing for accuracy and generalization. We then tested models on the test dataset. Conformal prediction methods were employed to assess model reliability, providing statistically sound confidence intervals for predictions. Model performance was visualized using line plots such as accuracy curves, confusion matrices and histograms to interpret the results effectively and highlight areas for improvement.

4.2 Technical Methods:

- **Logit Regression:** Traditional statistical methods for medical image classification, implemented using libraries such as OpenCV or Sci-Kit Learn.This model serves as a baseline for comparison with advanced models.
- **Inception V3**: A deep convolutional neural network designed for image classification and object detection, known for its efficient architecture that reduces computational complexity. It utilizes techniques like factorized convolutions, batch normalization, and auxiliary classifiers to improve accuracy and reduce overfitting. We use the pre-trained model by initializing it with the trainable parameters pre-trained on the ImageNet dataset, then we start training.
- **FixCaps:** An advanced model incorporating a Capsule Network and CBAM, which has shown world-leading performance. Implementation will follow the model's published code.

For the Logit Regression, Inception V3 model we trained for 20 epochs, and for the FixCaps model we trained for 120 epochs. In short, we apply existing model architectures and train them after selecting appropriate hyperparameters. Conformal Prediction (CP) provided us with prediction sets specifically for multi-class classification. We utilized CP for Logit Regression, Inception V3, and FixCaps. The reasons for not using Group Balanced CP(GBCP) will also be discussed in the challenges section, primarily due to insufficient image numbers in each class.

5 Results

5.1 Logit Regression

The Logit Regression model was employed in this project as a baseline for classifying skin lesions into seven categories using the HAM10000 dataset(2). Logit regression, a statistical model for classification tasks, predicts probabilities by modeling the relationship between independent and dependent variables through a logistic function(7). For this project, we used the multinomial logit model, which extends logistic regression to handle multiple output classes effectively and provides probabilistic predictions crucial for interpretability. While limited in handling high-dimensional

image data, the model's simplicity, interpretability, and use of a single softmax layer made it a useful benchmark for comparison against more advanced models like Inception V3 and ResNets, which also use a softmax layer as their final layer.

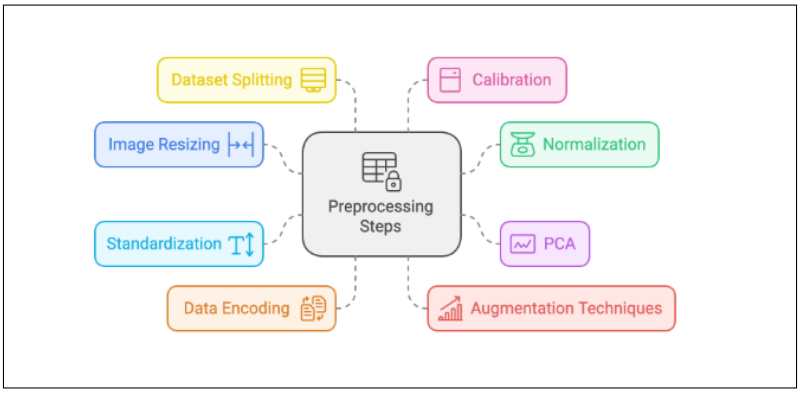

Figure 8: *Preprocessing steps for Logit Regression*

As seen in the above figure 8, preprocessing steps were critical for preparing the dataset. Images were resized to smaller dimensions (32x32 pixels) to reduce computational complexity. The pixel values were normalized to ensure consistent input scaling, and PCA was applied to reduce the feature set to 200 principal components, retaining most of the variance while minimizing complexity. Additional preprocessing included standardization and numerical encoding of data, as well as augmentation techniques such as rotation, flipping, and cropping to address class imbalances. The dataset was divided into training, validation, and testing subsets, and calibration was applied to enable conformal prediction.

To optimize the logit regression model, regularization was employed to minimize overfitting, and the lbfgs solver was selected for its efficiency in handling multinomial settings and faster convergence. Despite these optimizations, challenges such as computational complexity and convergence issues were encountered, requiring careful hyperparameter tuning. We observed that training on a PC required approximately 4 hours per epoch. To enhance efficiency, we transitioned to a GPU cluster, reducing training time to about 1 minute per epoch using a 10GB slice of NVIDIA Ampere A100-80GB GPUs. To get the optimal training, we chose the learning rate to be 1e-5 via the learning rate range test and deployed penalization to relieve the problem of overfitting. For details of the hyperparameter tuning on the GPU cluster, please refer to the Appendix 1.

Type	Precision	Recall	F1	Accuracy
akiec	0.1765	0.07	0.1	
bcc	0.3803	0.29	0.329	
bkl	0.3061	0.069	0.113	
df	0.0952	0.091	0.093	
mel	0.2895	0.064	0.105	
nv	0.6776	0.914	0.778	
vasc	0.2174	0.429	0.288	
Total:				0.5989

Figure 9: *Classification Report for Logit Regression*

The line plots below 10, show that the logit regression model achieved an overall accuracy of approximately 35%, 59.89% and 67.27% on the train, test, and validation datasets respectively. Class-level performance, evaluated using precision, recall, and F1 scores, revealed significant variability (as seen in figure9). The model excelled in identifying 'nv' lesions, with an F1 score of 0.778, precision of 0.6776, and recall of 0.914. However, it struggled with other lesion types, such as 'akiec' and 'df,' yielding F1 scores as low as 0.1 and 0.093, respectively. These disparities highlight its limitations in generalizing to underrepresented or complex classes. A confusion matrix further illuminated misclassifications, and conformal prediction estimated an average prediction set size of 3.8, which is over the half of the overall size of all classes(7), providing the potentially fragile model confidence.

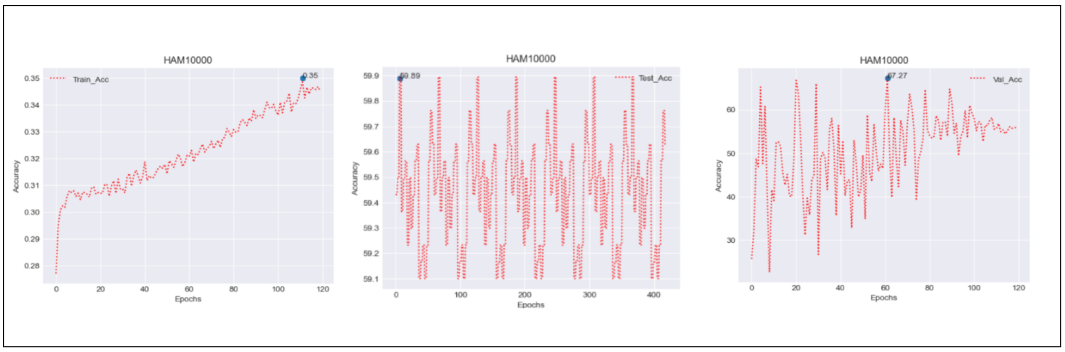

Figure 10: *Accuracy plots for Logit Regression*

While the multinomial logit model provided valuable probabilistic predictions and interpretability, its inability to capture non-linear relationships and spatial features limited its effectiveness for medical image classification. Computational constraints, including large dataset memory requirements and crashes on multiple systems, further restricted the model's potential. Although regularization and hyperparameter tuning improved the results, this model's inherent simplicity made it unsuitable for capturing the intricacies of image data. Deep learning models, such as Inception V3, were explored as the next step to address these challenges and improve performance.

5.2 Inception V3 model

To enhance the classification performance, we adopted the Inception V3 model, which is a deep convolutional neural network (CNN) architecture designed specifically for image classification and feature extraction tasks (8). In this project, we employed this model to classify skin lesions into categories and to compare with the baseline model Logit Regression. This model was selected due to the following advantages:

- Stronger Representational Power: Compared to other deep learning models such as VGG, the Inception V3 model is a deeper and larger model which has the ability of capturing more complex features in the images.
- Multi-scale feature Extraction: Inception V3 model can simultaneously extract features at different scales, improving the ability to classify features.
- Stable and Easy Training Process: The training process includes techniques such as Batch Normalization and Auxiliary Classification, which stabilized the training process.

The preprocessing steps are the same with those for the Logit Regression deployed on the cluster. We use a pre-trained Inception V3 model originally trained on the ImageNet dataset and fine-tuned it on the HAM 10000 dataset. After 20 epochs (approximately 8 minutes per epoch) of training and with a 0.003 learning rate which was selected based on the learning rate tuning.

As the result plots below 11, the Inception V3 model demonstrated significant improvements in classification accuracy, achieving an overall training accuracy of 98%, a peak validation accuracy of 88.88%, and a test accuracy of 81.34%. Despite its superiority over the logit regression, the pre-trained model overfits the data and it is probably due to the excessively large number of trainable parameters compared to relatively small amounts of data.

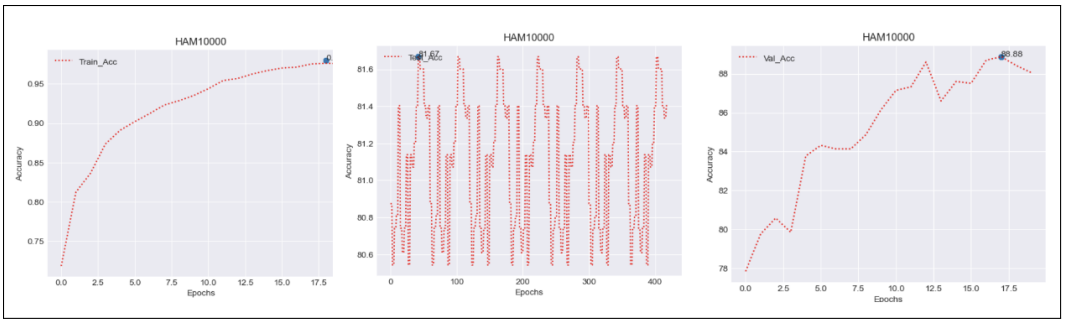

Figure 11: *Inception V3 Accuracy Plots*

Type	Precision	Recall	F1	Accuracy
akiec	0.6667	0.465	0.548	
bcc	0.8243	0.656	0.731	
bkl	0.7831	0.682	0.729	
df	0.9333	0.318	0.475	
mel	0.6741	0.532	0.595	
nv	0.8362	0.961	0.894	
vasc	0.9167	0.629	0.746	
Total:				0.8134

Figure 12: *Inception V3 Classification Report*

Class-level performance metrics revealed that Inception V3 consistently outperformed Logit Regression, particularly in identifying complex lesion types. Precision, recall, and F1-score for each class are higher than the Logit Regression model. The classification report in Figure 12 highlights the improvement of classification confidence across all categories. However, we observed that the Inception V3 model's performance on certain individual classes, such as 'akiec' and 'df', was still relatively low, with F1 scores of 0.6667 and 0.6741. We infer that this result may be due to the rarity of these classes in the test dataset, making it challenging for the model to effectively learn their features.

Model	Params(M)	FLOPs(G)	FPS
InceptionV3	21.8	5.75	142.97

Figure 13: *Inception V3 Model Parameters*

We used confusion matrix to understand the model's predictions to the actual labels. This helped us assess how effectively the model differentiates across classes in addition to its overall accuracy. The model's complexity and possible computational needs are indicated by the parameter count, which refers to the number of learnable weights in the model. The amount of work the model must perform to process an input is shown by FLOPs (Floating Point Operations), a measure of the computing effort needed for inference. FPS (Frames Per Second) gives the model's speed, or how rapidly it can generate predictions.

Based on these metrics, Inception V3 model, (as seen in figure 13), has a high efficiency, with a parameter count of 21.8M and a computational cost of 5.75 GFLOPs. These features enabled the model to achieve high throughput and fast inference, as evidenced by a frame per second (FPS) rate of 142.97. We found that the conformal prediction estimated an average prediction set size of 2.7.

5.3 FixCaps(3)

To further improve the performance for classifying skin lesions by using the HAM10000 dataset(2), we propose an improved Capsule Network model: FixCaps. Compared to the previously used Inception V3 model, FixCaps has the following advantages:

- Strong Feature Representation Capability: FixCaps can learn richer and more hierarchical visual features, and can better capture complex visual patterns compared to convolutional neural networks.
- Robust Classification Performance: FixCap's dynamic routing mechanism can better handle image transformations such as deformation, translation, and rotation, improving the model's generalization ability.
- High Parameter Efficiency: FixCaps significantly reduces the number of model parameters by introducing fixed-length capsules and a dynamic routing algorithm, improving computational efficiency.

We implemented the FixCaps architecture as mentioned in the literature review. The methodology and parameter settings are as follows:

- The initial learning rate was set to 0.123.
- A cosine annealing strategy was used to gradually decrease the learning rate during training.
- The number of primary units is an intermediate layer parameter, but the details were not explored further.
- The output unit size was set to 16, resulting in a final output torch size of batch size is $[\text{batch size}, 7, 16, 1]$
- The model parameter is an interface set by the researchers to modify the order and quantity of layers within the network, but the specifics were not delved into.

Model	Params(M)	FLOPs(G)	FPS
FixCaps-128	0.46	5.99	119.8

Figure 14: *Model Parameters of FixCaps*

This model has a parameter count of 0.46M which is much lower than the Inception V3 model and a computational cost of 5.99 GFLOPs which is higher. These features enabled the model to achieve high throughput and fast inference, as evidenced by a frame per second (FPS) rate of 119.8 (See figure 14).

The preprocessing steps are the same as those used for the Logit Regression model and Inception V3 model. During the training process, we observed that the training loss reached the lowest point at epoch 60. Also because of the runtime limitation of 4 hours, we recorded the model parameters at this epoch for further training. The entire training process took 8 hours to complete 120 epochs.

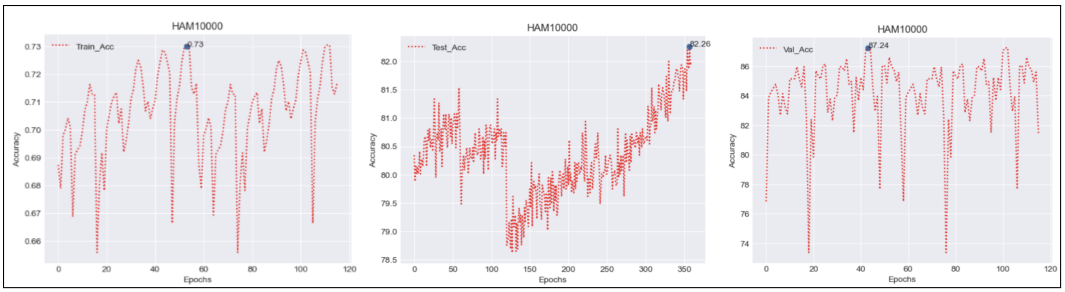

Figure 15: *Accuracy Plots for FixCaps*

After 120 epochs of training, the FixCaps model achieved a peak validation accuracy of 87.24% and a test accuracy of 82.26%, significantly outperforming the Inception V3 model 15.

We also assume that we will not lose any information gained from the previous training if we save the status of the model, the optimizer, and the scheduler. We split the training into two parts, each running for 4 hours, since the maximum runtime of the GPU cluster is 4 hours. After finishing the first half of the training (about 60 epoches), we save the information in a checkpoint and reload it at the beginning of the second half of the training.

Type	Precision	Recall	F1	Accuracy
akiec	0.4286	0.279	0.338	
bcc	0.664	0.892	0.761	
bkl	0.6727	0.862	0.756	
df	0.4667	0.318	0.378	
mel	0.9041	0.386	0.541	
nv	0.9018	0.96	0.93	
vasc	0.9	0.257	0.4	
Total:				0.8226

Figure 16: *Classification Report for FixCaps*

From the class-level performance matrix for the FixCaps model (See figure 16), it can be inferred that the model achieved the highest accuracy. The confusion matrix (can be referred to in Appendix 3) shows that the FixCaps model has high prediction accuracies on most classes, especially for the more common classes like 'nv', 'bkl', and 'mel'. We found that the conformal prediction estimated an average prediction set size of 1.86.

6 Model Comparison and Discussion

We compared the three models with the Conformal prediction techniques. It helps one gain a qualitative understanding of the prediction confidence of models. For example, a classifier uses its softmax scores diagnosing the image as "df", a kind of benign skin cancer, but what if the doctor wants to know how likely the image is to be classified as 'mel' or 'bcc' or other kinds of malignant cancer? Conformal prediction gives the prediction set that provably covers the true diagnosis with a given confidence level.

For the conformal predictions on all the three models trained on the GPU cluster, a method called Regularized Adaptive Prediction Sets (RAPS)(6) are implemented for producing prediction sets. RAPS is an improved version of Adaptive Prediction Sets (APS). Unlike the previously mentioned example of calculating the conformal scores, which is

$$S(X_{\text{test}}, y_{\text{test}}) = 1 - \text{Softmax}(X_{\text{test}})_{y_{\text{test}}}$$

APS utilizes the softmax outputs of other classes, not just the true class, and by integrating more information, it is able to avoid the problem of undercovering tricky cases and overcovering easy cases.

Its score function is calculated by greedily include the softmax scores of the classes until the true label is reached:

$$S(X_{\text{test}}, Y_{\text{test}}) = \Sigma_{j=1}^{k} \text{Softmax}(X_{\text{test}})_{\pi_j}, Y_{\text{test}} = \pi_k$$

in which $\{\pi_1, \pi_2, ..., \pi_K\}$ is a permutation of $\{1, 2, ..., K\}$ that sorts the softmax scores of X_{test} from most likely to least likely.

RAPS achieves better performance (a smaller prediction set on average with the same confidence level) because it places penalization on overly large sets by introducing two tunable parameters k_{reg} and λ. Every class beyond the k_{reg} most likely classes is subject to a constant penalty λ.

The two hyper parameters k_{reg} and λ are chosen by cross validation with the optimization goal as minimizing the average size of prediction sets. For detailed deployment, please refer to Appendix 2. We split the test dataset into the calibration and the validation set randomly with the calibration set including roughly 1000 images and 500 images left for the validation set, the hyper-parameters might change due to a different way of splitting. Please refer to the Appendix 2 on why using at least 1000 images for calibration.

6.1 Correctness Check

First correctness checks are performed to guarantee the reasonableness of any further examinations. We empirically checked the coverages of 1000 independent trails to see if they satisfies the marginal coverage theorem.

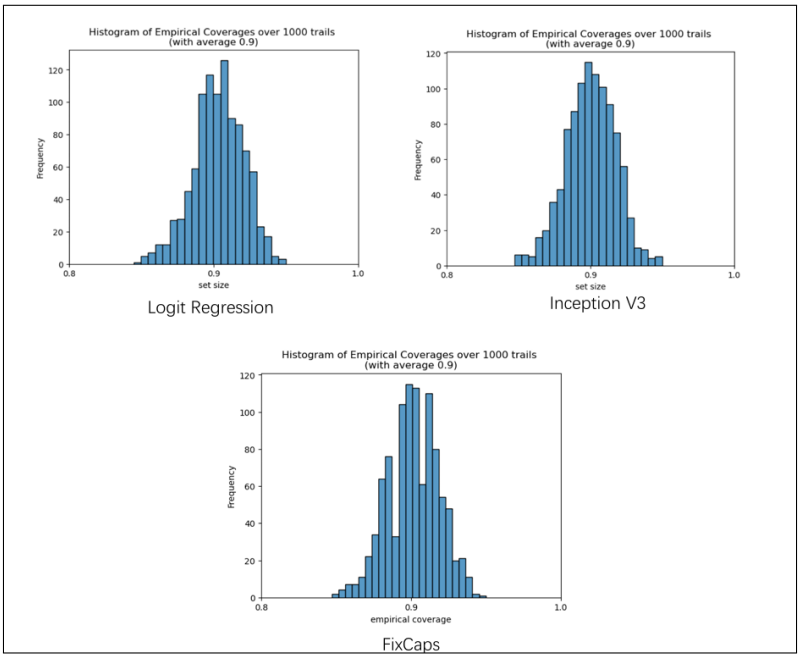

Figure 17: *Correctness Check of Conformal Prediction*

Experiment results show (See fig 17) that the empirical coverages in average are 0.9, which is exactly $1 - \alpha$ when $\alpha = 0.1$.

6.2 Adaptivity Check

Distribution of Prediction Set Size

The second step is to evaluate the adaptivity. Naturally, it is expected that the size of the prediction sets reflect the model's confidence, or the difficulty of making predictions on a given validation case. The tricker the case, the larger the sets. But this property is not guaranteed by the marginal coverage theorem, while it is important in practical deployments. The larger the average set size, the less precise the conformal procedure. Since the conformal procedure, confidence levels, and validation

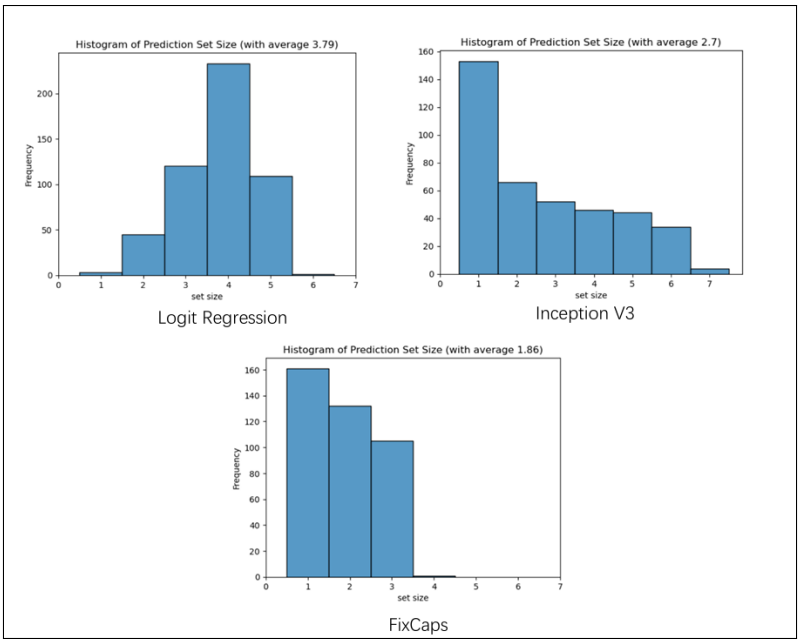

Figure 18: *Distribution of the Prediction Set Size*

data are all the same, the average set sizes are comparable among the three classifiers. FixCaps has the smallest average set size (which is 1.86), the average set size of Inception V3 is in the middle (which is 2.7) and Logit Regression has the largest (which is 3.79). See fig 18. The results show the confidence difference among the three models. The two deep learning models outperform the logit regression again from this perspective.

The generally desirable distribution of the prediction set size should have a wide spread, and the three histograms in fig 18 roughly satisfy the requirement. We define a case being "difficult" for a model as its prediction set size being large. Since a dataset consists of cases of varying difficulties, and we have no quantitative metric to define the difficulty, the conclusions derived are that an easy case from the perspective of one classifier might be difficult for the other classifier, and that the classifier's prediction patterns vary. The logit regression is "eclectic" so it always chooses the middle way — including 2 to 5 classes within a prediction set in most cases. The other two classifiers, on the other hand, only include one class for easy cases within the prediction set and become less certain and include more when encountering difficult ones. The figures above also show that they have confidence in predicting most cases and are only uncertain on a few cases.

After the above qualitative analysis, it is reasonable to doubt whether the logit regression acquires the ability to classify skin cancers or not since it adopts a strategy similar to eclecticism.

Conditional Coverage
A quantitative metric for adaptivity is related to the conditional coverage:

$$P\left(Y_{\text{test}} \in C(X_{\text{test}})|X_{\text{test}}\right) \geq 1-\alpha.$$

For every specific input case, the conditional coverage guarantees $1-\alpha$ coverage, which is stronger than the marginal coverage that the conformal prediction is guaranteed to achieve. So it is necessary to check how close the procedures are from the conditional coverage. The size-stratified coverage metric (SSC) is a commonly used metric for measuring the degree of achieving the conditional coverage.

$$\text{SSC} = \min_{g\in 1,\ldots,G} \frac{1}{|\mathcal{I}_g|} 1\{Y_i^{\text{val}} \in C(X_i^{\text{val}})\}$$

By discretizing the cardinalities of prediction sets into G bins and then dividing observations into groups based on the set cardinalities of these cases that fall into which bins, the minimum across the empirical coverage rate within each group is defined as SSC. Since the cardinality of a prediction set indicates the difficulty level for a classifier to predict the corresponding case, the closer SSC to the confidence level $1 - \alpha$, the closer the classifier to achieving the conditional coverage across different difficulty levels. SSC indicates whether or not a classifier tends to undercover hard subgroups and overcover easy ones. Table 1 shows the experiment results based on a random split of the test dataset, so they are only for a reference to show that Logit Regression is the furthest one away from achieving the conditional coverage than the other two deep learning models. (The confidence level is $1 - \alpha = 0.9$)

Table 1: SSC metric of the three models

Model	Logit	Inception V3	FixCaps
SSC	0.5	0.83	0.92

7 Challenges

The training process was challenging due to limitations in computing power. Initial training on a PC proved insufficient due to its lack of processing capacity to handle large datasets and complex models efficiently, necessitating a shift to GPUs.

While GPUs enabled faster training times, their maximum runtime of 4 hours per session set by the administrator was insufficient to adequately train sophisticated models like FixCaps, which need at least 8 hours to train effectively. We also started with Resnet implementation but decided not to go ahead with it and include in results due to limits on time and computing resources.

The dataset was also imbalanced and hence required the application of data augmentation techniques to ensure equitable representation across classes for improving model performance and fairness.

There were challenges to implement GBCP as the calibration set would be insufficient to ensure effective performance of the model as discussed above.

8 Future scope

Moving forward, our efforts would be to expand the dataset to include more diverse and underrepresented samples, optimizing model architectures for fairness and efficiency, and validating these models in real-world clinical scenarios. Additionally, we could extend our existing work on Conformal Prediction (CP) to Group-Balanced Conformal Prediction (GBCP). Implementing GBCP would address challenges like subgroup data sparsity and computational feasibility, further enhancing the fairness and reliability of our models.

Another promising direction is the implementation of ResNet architectures, which was not possible due to time constraints. ResNet's ability to capture intricate spatial features makes it a strong candidate for evaluating its performance on the inclusive HAM10000 dataset.

Additionally, improving the second-half training efficiency of FixCaps through techniques like dynamic learning rate adjustments or optimized training strategies could reduce computational demands while maintaining accuracy. These efforts will help refine the fairness, accuracy, and practicality of medical AI models.

9 Conclusion

This project has made a significant impact by addressing a critical gap in medical AI: the need for inclusive and equitable diagnostic models for skin cancer detection across diverse skin tones. By developing and evaluating multiple models on the HAM10000 dataset, we have advanced the state of skin lesion classification, contributing to more accurate and fair diagnostic tools that can aid early detection for all demographics. The inclusion of fairness metrics, such as Conformal Prediction

(CP), ensured balanced performance across lighter and darker skin tones, underscoring the ethical imperative of reducing healthcare disparities through AI.

All project goals were successfully achieved. We developed and rigorously evaluated three models—Logit Regression, Inception V3, and FixCaps—assessing their performance on classification accuracy, precision, recall, and subgroup fairness. Each model served a specific purpose in this progression, from establishing a baseline to achieving state-of-the-art performance. The models were compared in depth, and FixCaps emerged as the most effective approach. With its peak test accuracy of 82.26%, FixCaps demonstrated strong feature representation capabilities, efficient parameter utilization, and robustness in handling complex medical image data. It outperformed Inception V3 and Logit Regression, particularly in managing underrepresented classes and maintaining consistent performance across all categories. While Inception V3 offered significant improvements over the baseline, FixCap's dynamic routing mechanism and reduced parameter count solidified its superiority.

This work highlights the transformative potential of advanced deep learning architectures in healthcare. By prioritizing fairness and inclusivity, we have laid the groundwork for creating equitable diagnostic systems that address both technical and ethical challenges. The outcomes not only contribute to the field of medical AI but also establish a foundation for future research and clinical applications.

References

[1] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A. (2018) Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). Available at: `https://arxiv.org/abs/1902.03368`.

[2] Tschandl, P., Rosendahl, C., & Kittler, H. (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161. doi: `https://arxiv.org/abs/1803.10417`.

[3] Li, M., Chen, H., Peng, J., Li, X., Zhang, Y., & Lin, J. (2022) FixCaps: An Improved Capsules Network for Diagnosis of Skin Cancer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11), 12165–12173. Available at: `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9791221`.

[4] Lu, C., Lemay, A., Chang, K., Hobel, K., & Kalpathy-Cramer, J. (2022) Fair Conformal Predictors for Applications in Medical Imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11), 12008–12016. doi: `https://doi.org/10.1609/aaai.v36i11.21459`.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. doi: `https://doi.org/10.1109/CVPR.2016.90`.

[6] Angelopoulos, A. N., Bates, S. (2021) A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint*, arXiv:2107.07511. Available at: `https://arxiv.org/abs/2107.07511`.

[7] Yang, H., Zhong, Z., Wang, R., and Liu, H. (2021). Data Augmentation in Logit Space for Medical Image Classification with Limited Training Data. *ResearchGate*. Retrieved from `https://www.researchgate.net/publication/354783735_Data_Augmentation_in_Logit_Space_for_Medical_Image_Classification_with_Limited_Training_Data`.

[8] Yuhang, P., Liu, J., and Yang, X. (2023). Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases. *PMC*. Retrieved from `https://pmc.ncbi.nlm.nih.gov/articles/PMC9975334/`.

[9] STA 221 Project. *GitHub repository*. Retrieved from `https://github.com/Yehongqiu-lab/STA-221-Project`.

[10] HAM10000 Dataset. Available: `https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000`

A Appendix 1: Hyperparameter Tuning Details for Model Training

A.1 Logit Regression

Determine the learning rate by the learning rate range test. Specifically, the learning rate is set to increase at an exponential rate with $\gamma = 1.1$:

$$\text{lr\_nw} = \text{lr} \times \gamma^{\text{epoch}}$$

Every time a learning rate is updated, calculate the model loss on one batch of data after performing one gradient descent step. Choose the largest learning rate that makes the loss rapidly decrease.

$$\text{lr} = \text{1e-5}$$

Figure 19: *Learning Rate Range*

Determine the penalty term parameter `weight_decay` by training models with different penalized terms for 5 epochs and then validating them on the validation set:

$$\begin{aligned}
&\text{weight_decay} = 0.0001: \quad \text{val_loss} = 1.4625, \quad \text{val_accuracy} = 0.5059 \\
&\text{weight_decay} = 0.001: \quad \text{val_loss} = 1.4369, \quad \text{val_accuracy} = 0.5160 \\
&\text{weight_decay} = 0.01: \quad \text{val_loss} = 1.4660, \quad \text{val_accuracy} = 0.5105 \\
&\text{weight_decay} = 0.1: \quad \text{val_loss} = 1.4961, \quad \text{val_accuracy} = 0.4859
\end{aligned}$$

A.2 Inception V3

For Inception V3, the learning rate is set as 1e-3. The penalized term `weight_decay` is set in the same way as for the logit regression.

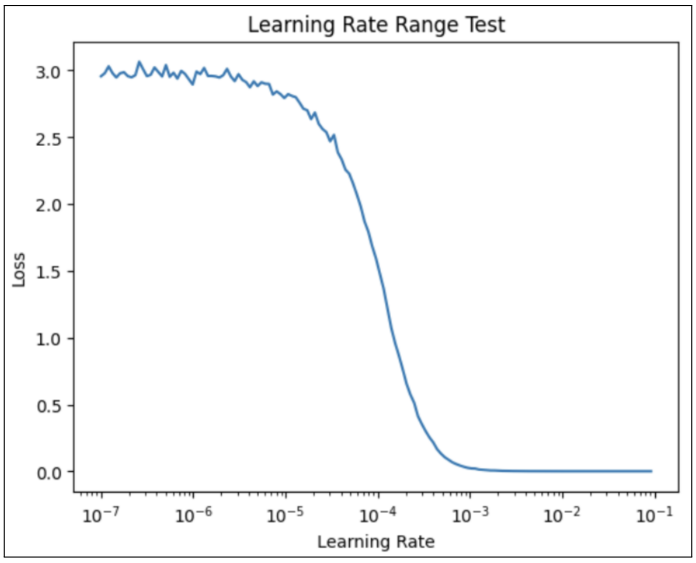

Figure 20: *Learning Rate Range*

B Appendix 2: Technical Details about Conformal Prediction

B.1 Regularized Adaptive Prediction Sets (RAPS)

The two hyper parameters k_{reg} and λ are chosen by cross validation with the optimization goal as minimizing the average size of prediction sets. The chosen parameters vary model from model. In practical deployments, RAPS algorithm also introduces two Boolean variables for users to slightly change the behavior of the algorithm. The options are whether to allow prediction set size to be 0 or not, and whether to introduce a small random number into the conformal scores calculation to help break the potential tie. We set both of the two Booleans as True. Since we split the test dataset used for testing the three models before into the calibration and the validation set randomly with the calibration set including roughly 1000 images and 500 images left for the validation set, the parameters shown in the Table 2 might change due to a different way of splitting. The k_{reg} is set to be 0, which means that the RAPS implemented here is actually APS, so the λ is not necessary for Logit Regression.

Table 2: Chosen Hyper-Parameters for Deploying RAPS

Model	k_{reg}	λ
Logit	0	N/A
Inception V3	2	5e-8
FixCaps	1	0.05

B.2 Size of the Calibration set

The size of the calibration set affects conformal prediction. The key idea is that the coverage of conformal prediction conditionally on the calibration set is a random variable.

$$P\left(Y_{\text{test}} \in C(X_{\text{test}}) \mid \{(X_i, Y_i)\}_{i=1}^{n}\right) \sim \text{Beta}(n+1-l, l),$$

$$\text{in which } l = \lfloor (n+1)\alpha \rfloor.$$

When choosing $n = 1000$ and $\alpha = 0.1$, the coverage is between 0.88 and 0.92.

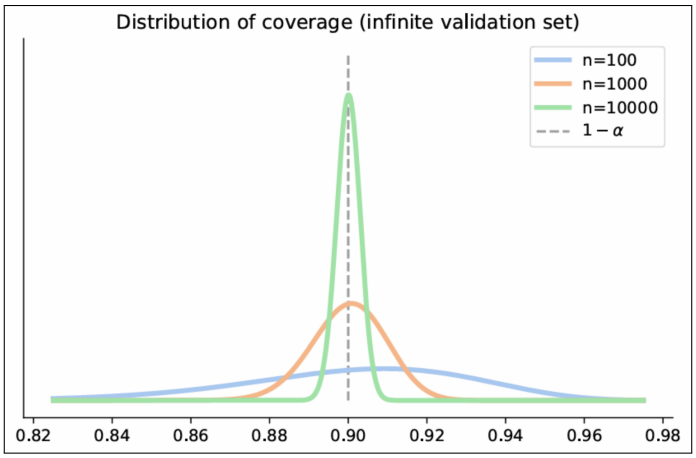

Figure 21: *Calibration Set for Conformal Prediction*

C Appendix 3: Additional Results

C.1 Logit Regression

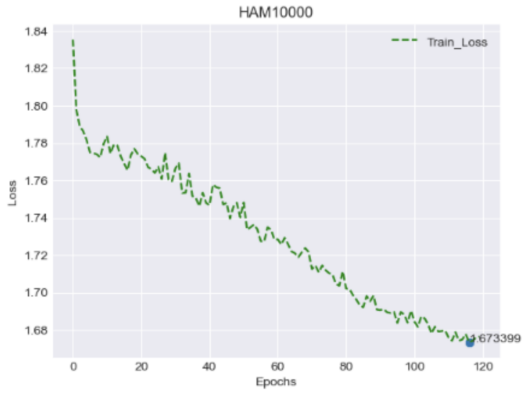

Figure 22: *Logit Regression Training Loss*

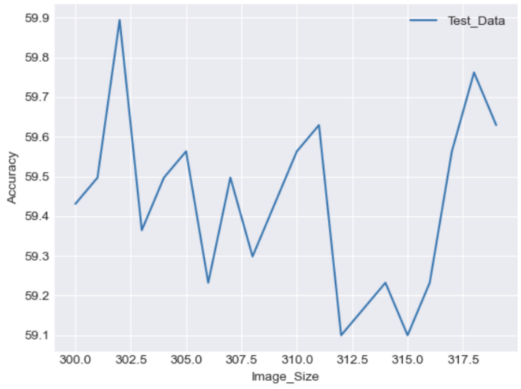

Figure 23: *Logit Regression Test Data*

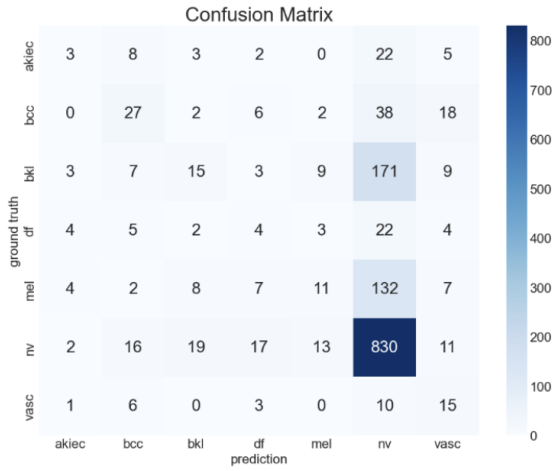

Figure 24: *Confusion Matrix for Logit Regression*

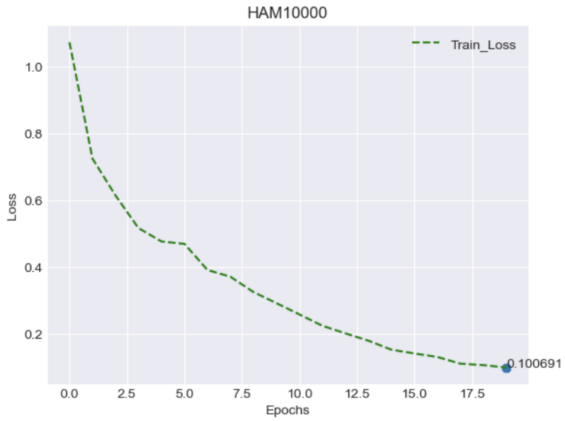

Figure 25: *Inception V3 Training Loss*

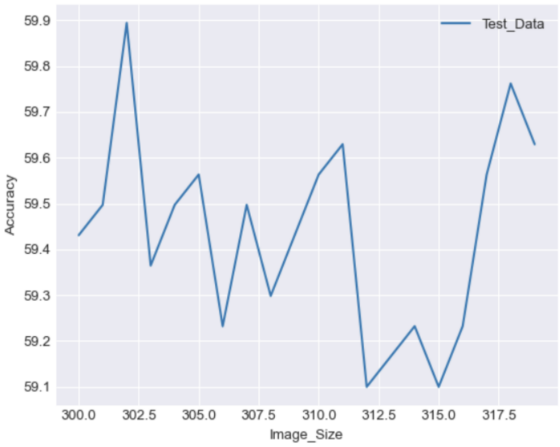

Figure 26: *Inception V3 Test Data*

Figure 27: *Confusion Matrix for Inception V3 Model*

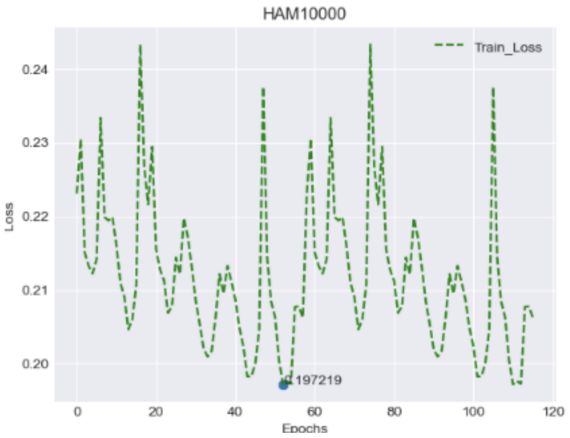

Figure 28: *FixCaps Training Loss*

Figure 29: *FixCaps Test Data*

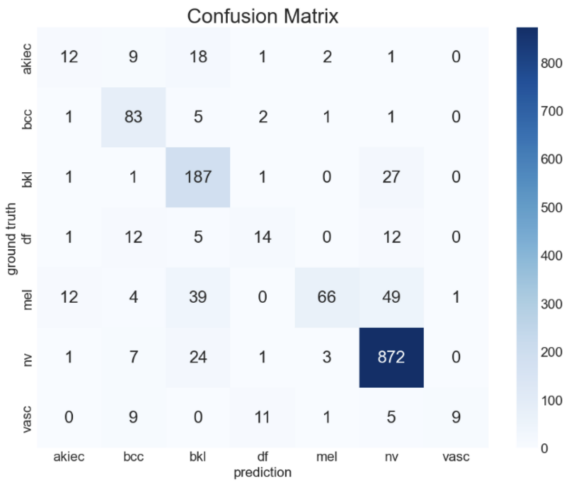

Figure 30: *Confusion Matrix for FixCaps Model*