# Beyond the Surface:
# Advancing Skin Cancer Detection Across All Tones

ChiChun Chen, Savali Sandip Deshmukh, Shivani Sanjay Suryawanshi, Sriharshini D, and Yehong Qiu

## 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 Group-Balanced Conformal Prediction (GBCP), a distribution-free uncertainty quantification technique on a model's confidence of delivering correct outcomes. We will compare Convolutional Neural Networks (CNNs) with other supervised models to determine the best-suited architecture for this task. 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

### Research Questions

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

#### RQ1. Classification Effectiveness Across Skin Tones

Which deep 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.

### RQ2. Model Comparison in Supervised Learning

How do CNNs 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 deep learning models within a supervised learning context, helping identify the most effective architecture for skin cancer detection across diverse skin tones.

### 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 GBCP and also some other Epistemic uncertainty measures 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 [5] dataset 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. This categorization is vital for our primary prediction goal, which is to classify skin lesions accurately.

We will source the HAM10000 dataset from publicly available platforms like Kaggle and will segment it based on skin tone categories to ensure representation across lighter and darker skin tones. Preprocessing the data will involve several steps: resizing images to a consistent dimension, normalizing pixel values to enhance model performance, and augmenting the dataset through techniques like rotation, flipping, and zooming to increase the variability and robustness of the model. Metadata will be cleaned to handle missing values and ensure consistency. 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.

# 3 Literature Review

## FixCaps: An Improved Capsules Network for Diagnosis of Skin Cancer

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, 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. It also reduces computational costs compared to previous methods with fewer parameters, making it efficient for clinical applications.

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.

## Fair Conformal Predictors for Applications in Medical Imaging

We chose the following paper[3] because we want to evaluate the reliability of DL models on sub-group accuracy, and Conformal Prediction (CP) 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, $\alpha$, 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.

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.

# 4 Methodology

## 4.1 Workflow

Our workflow involves several steps: First, we conduct Exploratory Data Analysis on the chosen dataset. Then, Data Augmentation will be implemented if the data is imbalanced, including techniques such as rotation, flipping, scaling, cropping, color jittering, noise injection, and advanced methods like MixUp and GANs.Next, we will split the augmented dataset into training, validation, and testing sets. After data-pre-processing, we will train machine learning models on the training dataset. Then, we will construct Group-Balanced Conformal Prediction on each ML classifier to evaluate model reliability. Lastly we will use heatmaps to visualize model performance and interpret results.

## 4.2 Technical Methods:

- **Logit/Probit Regression**: Traditional statistical methods for medical image classification, implemented using libraries such as OpenCV or Sci-Kit Learn.
- **Convolutional Neural Network (CNN)**: A classic deep learning model for image classification, to be used as a benchmark for other advanced models. Implementation will reference the textbook "Deep Learning" by Ian Goodfellow et al. and the Pytorch library.
- **ResNet**: An evolution of CNN that solves training efficiency problems in deep networks. We will use a pre-trained ResNet model from ImageNet and implement it on our dataset.
- **FixCaps**: An advanced model incorporating a Capsule Network and CBAM, which has shown world-leading performance. Implementation will follow the model's published code.
- **Conformal Prediction (CP)**: Provides confidence sets specifically for multi-class classification.

# 5 Conclusion

For research question 1, we expect that FixCaps will classify skin cancer lesions with the highest accuracy across a range of skin tones. FixCaps has demonstrated its potential as the most accurate model for complex lesion images by achieving up to 96.49% accuracy on the HAM10000 dataset. [1].We may see a minor decline, though, as a result of extra fairness modifications made to guarantee performance across a range of skin tones. FixCaps should have accuracy between 94 and 95%. Given its capacity to manage deep structures, ResNet can also work effectively whereas simple CNNs and traditional models like logistic/probit regression are predicted to deliver lower baselines at 85-88% and 70-75% accuracy, respectively. We may compare these models to see which architecture responds better to changes in skin tone.

For research question 2, we expect CNNs and ResNet to perform better than logistic/probit regression in terms of model comparison in supervised learning since they are able to capture and learn from high-dimensional image data. We consider FixCaps to be a benchmark because of its sophisticated structure that preserves spatial data with group convolution while ResNet's depth gives it an advantage over simple CNNs by addressing training and gradient concerns.

Common metrics for assessing classifier accuracy include sensitivity, specificity, and subgroup accuracy (lighter vs. darker skin tones). Conformal prediction techniques, especially Group-Balanced Conformal Prediction (GBCP), can enhance fairness across skin tones by quantifying uncertainty and providing fair prediction intervals. We anticipate approximately 90% coverage, with per-group variation within ±3%.[3] Initially, accuracy may vary across skin tones, but with fairness tuning, accuracy should converge to 88–92% across demographics. Conservative prediction intervals may be required for underrepresented skin tones and larger sets for ambiguous cases. [2]

The cardinality of prediction sets measures the confidence of ML models. Expected prediction set sizes range from 1.4 to 1.8 classes (60% single-class, 30% two-class, and 10% three or more classes). These adjustments highlight the trade-off between accuracy and fairness, essential for trustworthy medical AI, while reflecting both the potential and challenges discussed in the literature.

We aim to evaluate classification accuracy for the HAM10000 dataset's multi-class classification across its seven categories of skin lesions, particularly with methods successful in managing imbalanced datasets and multi-class scenarios. By refining CNN architectures and employing advanced models like FixCaps, we may get a better understanding how these models contribute to fairness across different skin tones.

# References

[1] M. Li, H. Chen, J. Peng, X. Li, Y. Zhang, and J. Lin, *FixCaps: An Improved Capsules Network for Diagnosis of Skin Cancer*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 11, pp. 12165-12173, 2022. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9791221

[2] Dembczynski, K., Nadeem, A. (2022). *Pitfalls of Conformal Predictions for Medical Image Classification*. In: G. M. G. Portela, R. Santos, R. G. R. Costa (Eds.), *Advances in Artificial Intelligence and Data Engineering*. Springer. DOI: https://doi.org/10.1007/978-3-031-44336-7_20

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

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

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