Age-specific Detection of Skin Cancer Using Tree-Based and Deep Learning Models

Tianhong Shen, Weiguo Jiang
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
{t54shen,w33jiang}@uwaterloo.ca

Abstract

Skin cancer is the abnormal development of skin cells and primarily occurs on skin that is overly exposed to the sun. Most skin cancer can be cured easily at its early stages. However, the initial development of it looks no different than a regular mole on the skin, and diagnosis of malignant skin cancer requires visiting the hospital. Therefore, people tend to overlook these small changes on their skin and potentially miss the opportunity for early treatment. In this paper, we present both a tree-based approach and a deep learning approach to predict whether a given dermoscopic image is benign or malignant. A classical machine learning model will be used as the benchmark for comparison. In particular, we are interested in fitting our models based on different age groups to assist potential patients better.

1 Introduction

According to Skin Cancer Foundation, skin cancer is "the most common cancer in the United States and worldwide", with "more people diagnosed with skin cancer each year in the U.S. than all other cancers combined" (Skin Cancer Foundation, 2023). At the same time, recent reports have called attention to a growing shortage of dermatologists per capita (Kimball and Resneck, 2008). Of all skin cancers, melanoma is the most dangerous category. Dermatology experts believe that in 2040, there will be half a million melanoma patients in the world, a 62% increase compared with that in 2018. Based on these statistics, skin cancer can be considered a global epidemic. To prevent the adverse consequences of skin cancer, dermatologists emphasize the importance of early treatment which can greatly improve the survival probability, especially for melanoma (See Figure 1) (Cassidy et al., 2022).

In this paper, we aim to detect the existence of malignant skin cancer given a dermoscopic image. The data are collected from the International Skin Imaging Collaboration (ISIC) gallery (ISIC Archive, b). Two approaches will be used, namely, extreme gradient boosting (XGBoost) which is tree-based, and convolutional neural network (CNN) that is a deep learning algorithm (Chen and Guestrin, 2016) (Krizhevsky et al., 2012). We will use the result from logistic regression as a benchmark for this binary classification problem. A positive label of 1 means that the image is malignant and a negative label of 0 represents benign. Additionally, to better serve different age groups, we used age information to tailor models, hoping to achieve a better performance than a generic classifier. Various metrics are employed to measure the performance of models against the benchmark.

2 Related Work

As with any kind of machine learning task involving images, pre-processing plays a big role in determining how good the performance of models can be. Dermatologists have their own set of rules

Figure 1: Benign vs Malignant. The left dermoscopic image is benign, and the right dermoscopic image is a malignant skin cancer, melanoma to be specific. It is hard to tell which one looks malignant based on appearance, and this is one of the main reasons why most people with malignant cancer tend to overlook it at the early stage.

for making a diagnosis. In this paper, two commonly used image-processing techniques for detecting skin cancers are used, namely ABCD and GLCM.

Asymmetry, Boundary irregularity, Color, Diameter (ABCD) method is the standard method for any dermatological applications. These four attributes of a skin lesion characterize the symptoms and aid in the detection of skin cancer. Asymmetry quantifies how two halves of an image differ; boundary irregularity measures the unevenness of the image; color is obtained by averaging the intensity of the three channels; and the diameter of the skin lesion is measured. The method of finding these parameters is termed the ABCD method (Monika et al., 2020).

Grey Level Co-occurrence Matrices (GLCM) is "one of the earliest methods for texture feature extraction", introduced by Haralick and his colleagues in 1973. Each entry (i, j) of the matrix represents the number of times a pixel of intensity i is adjacent to a pixel of intensity j (V. et al., 2012). This method is applied extensively in real-world problems, and it is still one of the most powerful feature extraction techniques in the field of texture analysis.

International Skin Imaging Collaboration (ISIC) hosted challenges over the years. Most of them focus on classifying the categories of skin cancer, such as melanoma, basal cell carcinoma (BCC), and Squamous cell carcinoma (SCC) (ISIC Archive, a). This led to the publication of several papers including the classification of malignant skin cancer into 8 different categories by M. Krishna Monika, et al.; on ISIC 2019 Challenge dataset, skin lesion detection towards melanoma by Noel C. F. Codella, et al.; on ISIC 2017 Challenge dataset, and B. Cassidy, et al created models to distinguish melanoma or non-melanoma (V. et al., 2012; Codella et al., 2017; Cassidy et al., 2022). Though our research on papers may not be exhaustive. However, after searching on common publication websites like arXiv, Google Scholar, and ResearchGate, we found that no work attempted to give a generic prediction on whether the dermoscopic scan is benign or malignant, let alone specializing with respect to age.

3 Preliminaries

3.1 Data Collection

The dataset used in this paper is collected from ISIC's gallery. All images are RGB with 3 channels and in jpg format. However, not all of them have the same dimension (i.e., length and width). All images are manually labeled with many attributes, including sex, benign or malignant, and approximate age. Several filters are available to assist researchers in retrieving customized data (ISIC Archive, b). Since we are specializing in detecting malignant skin cancers with respect to age, we mainly utilize the BENIGN OR MALIGNANT filter and APPROXIMATE AGE filter for data retrieval.

However, there are far more benign images than malignant images in the gallery. If we sample predominantly benign images, then it will introduce a considerable amount of dataset bias. To mitigate this issue, we will randomly select images from the gallery, with approximately 50% of benign and 50% of malignant as our training dataset. For the test dataset, we will keep the same

50/50 split of benign and malignant images to challenge our models.

We collect data based on age groups, namely 35-40, 50-55, 65-70, and 80-85. For each age group, 500 images are collected. Following the 50/50 split above, this means 250 benign images and 250 malignant images.

3.2 Data Pre-processing

Firstly, we try to remove bias and noise, such as duplicate and incomplete images. After removing them, we can minimize the impact of noise on model accuracy. Next, the main feature engineering methods we will use are ABCD and GCLM. They are used to extract numeric features from images.

To better illustrate the difference in skin cancer across age groups, we separate our dataset based on age groups as described in the data collection section above and develop separate models.

As mentioned previously, the dimension of each image varies which violates the model assumption of homogeneous input size. Therefore, we also applied downsampling to each image so that all images are kept at a fixed size.

3.3 Methodology

In general, we will apply XGBoost and CNN to these images. Meanwhile, we will use the result from logistic regression as the benchmark, so that we can make comparisons between models.

Since logistic regression is the benchmark, we choose not to dedicate separate models based on age models and train it directly on the entire dataset instead. For both XGBoost and CNN, we will develop four separate models for each age group and an overall model which trains on all age groups.

3.3.1 Logistic Regression (Benchmark)

Logistic regression is the fundamental model in dealing with binary classification. Packages can be applied directly to find the optimized parameters. However, due to its inherent inability to extract patterns and interpret features from raw images, logistic regression performs poorly on high-dimensional, hierarchical data.

3.3.2 Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting, also known as XGBoost, is one of the most powerful models in machine learning. It was invented by T. Chen and his colleagues in 2016, and it showed a strong performance on tabulated and structured data (Chen and Guestrin, 2016). In this paper, we have a high expectation on it.

3.3.3 Convolutional Neural Network (CNN)

Convolutional neural network (CNN) is an advanced image recognition method in machine learning, introduced by A. Krizhevsky, et al in 2012 (Krizhevsky et al., 2012). CNN can be directly applied to raw images, which is different from the two models we just mentioned. Till this day, it is still one of the most popular deep learning models for image processing tasks such as object detection, image classification, and facial recognition.

4 Evaluation

We will use two criteria to measure the performance of each model.

Accuracy is the most common criterion which measures the proportion of images that are correctly classified. Since the proportions of benign and malignant cases in the test set are similar, models cannot achieve a high accuracy score by always predicting either benign or malignant. Thus, accuracy serves as a representative metric to evaluate our models.

At the same time, to prevent the risk of misclassification, we will also give a confidence score which represents the probability of benign for each image. Thus, we can use cross entropy loss, also known as logarithmic loss, to measure the overall performance of each model. Its formula is

$$\textit{cross entropy loss} = \frac{1}{N} \times \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(p_{ij})$$

where N represents the total number of images, and M stands for the total number of classes (benign and malignant). $y_{ij} = 0$ if image i belongs to class j; otherwise, $y_{ij} = 1$. Moreover, p_{ij} refers to the probability of class j for image i.

5 Results

5.1 Logistic Regression

The benchmark of this experiment is generated using logistic regression. 80% of images are randomly selected to form the training set, and the remaining images are used as the test dataset to evaluate the performance of the model. As we mentioned earlier, each image is standardized to a fixed size, which is 64×64 with three channels. These features are flattened and used as the input of logistic regression, together with the ABCD and GLCM features, resulting in a total of 12,296 features.

Since the number of features is much larger than the number of samples, we apply the LASSO penalty in the linear operator to shrink the model. Additionally, the performance of logistic regression is sensitive to initializations. To reduce the instability, we repeat logistic regression for 100 times and evaluate each model on the test set. The accuracy and cross entropy loss benchmarks are computed as the average value of test accuracy and cross entropy loss, respectively. The average accuracy is 91.28%, and the average cross entropy loss is 0.2959.

5.2 XGBoost

The basic setup of XGBoost is similar to that of logistic regression, with features flattened. Sampling from observations and features introduce randomness to XGBoost models. To mitigate the uncertainty, we also run 100 times on XGBoost models and evaluate each model on the testset.

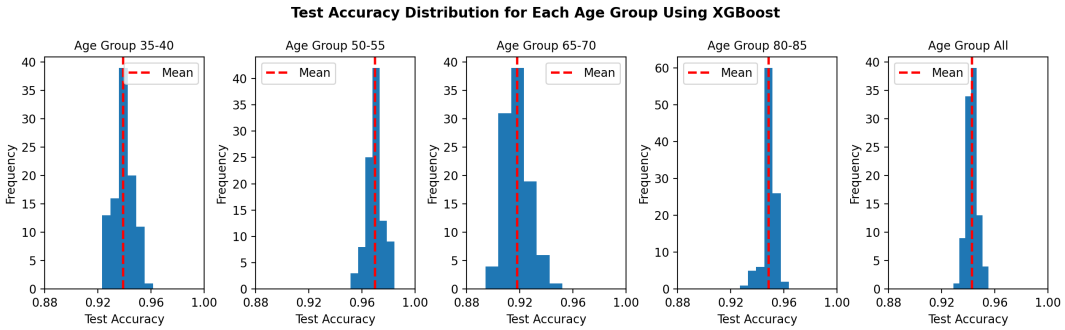

Figure 2: Histograms of Test Accuracy Distribution for Each Age Group Using XGBoost

A histogram of test accuracy for each age-specific model is presented in Figure 2. Compared to 94.3% average test accuracy of the overall model, the models for the age groups 50–55 and 80–85 tend to perform better, with mean accuracy 97% and 95%, respectively. In contrast, the model for the age group 65–70 performs less effectively with an accuracy 91.8%, as indicated by the red dashed lines. The model for the age interval 35-40 shows performance similar to that of the general model. Additionally, the bins for the overall model are more concentrated, indicating greater stability compared to the age-specific models. This may be due to the fact that the overall model benefits from

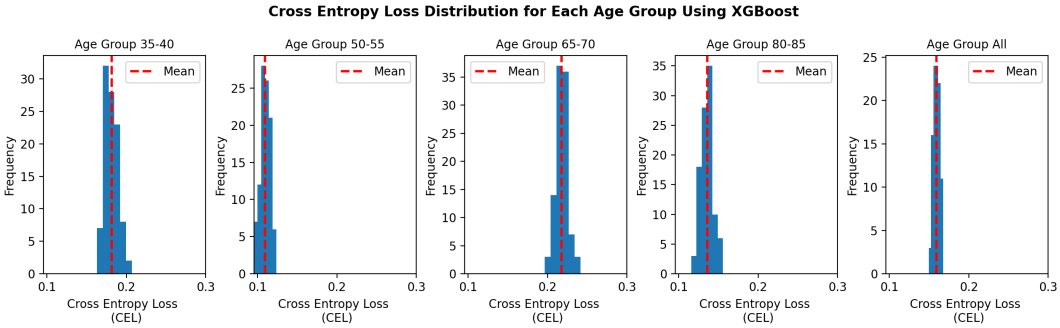

Figure 3: Histograms of Cross Entropy Loss Distribution for Each Age Group Using XGBoost

leveraging a larger and more diverse dataset.

Figure 3 demonstrates a histogram of cross entropy loss for each age-specific model. In comparison to 0.16 mean cross entropy loss of the overall model, the models for the age groups 50–55 and 80–85 usually achieve a lower cross entropy loss, with average loss 0.11 and 0.13, respectively. However, as indicated by the red dashed lines, the model for the age groups 35–40 and 65–70 show comparatively lower effectiveness, with losses of 0.18 and 0.22, respectively. Additionally, the distribution of the overall model's cross-entropy loss is more concentrated, suggesting a lower standard deviation compared to the age-specific models.

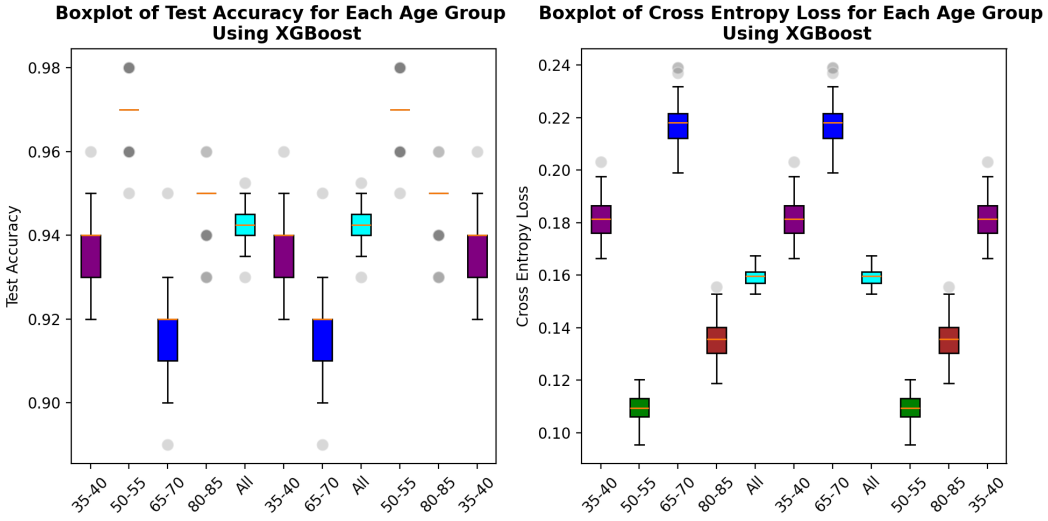

Figure 4: Boxplots of Test Accuracy (Left) and Cross Entropy Loss (Right) Distribution for Each Age Group Using XGBoost

Figure 4 provides a side-by-side comparison of distributions between different models. The boxplot of test accuracy on the left clearly demonstrates significant variation among the age groups. For example, for the overall model, the upper adjacent value is around 95%, and the lower adjacent value is approximately 93.5%. However, both the lower and upper adjacent value for age group 50-55 are 97%. Although there are many outliers, the general performance for age group 50-55 is still higher than that of the overall model. It means that the model for age group 50-55 extracts specific features from its training data and performs better, whereas the overall model primarily captures general patterns and leads to a lower test accuracy.

Next, the boxplot on the right reveals a significant variability in cross entropy loss across age groups. The general model shows no outliers, and its lower and upper adjacent values are relatively close. This observation aligns with the pattern identified in Figure 3, where the overall

model demonstrates greater stability in cross-entropy loss compared to the age-specific models. By comparison, models for age group 50-55 and 80-85 usually have a lower cross entropy loss than the overall model, indicating a higher confidence level.

Finally, a formal p-value test is conducted to determine whether all models have the same accuracy level. The test statistic is 401466 which follows an F distribution. With a p-value 0, and we conclude that the model performs differently across age groups. In other words, the XGBoost model offers potential improvements in skin cancer detection tailored to specific age groups.

5.3 CNN

The architecture of CNN is shown below in Figure 5.

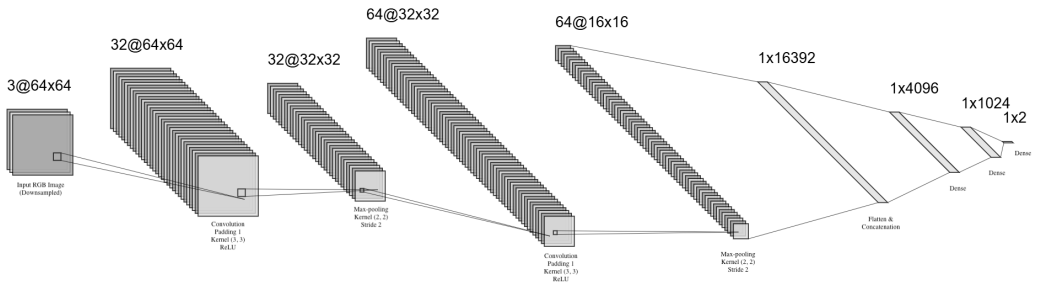

Figure 5: Architecture of CNN. Input is a downsampled 64×64 version of the original image. In the first convolution layer we have 32 kernels of size 2×2 and padding of 1 which helps preserve the dimensions. ReLU and max-pooling is applied which results in 32 feature maps of size 32×32. Second convolution layer has 64 kernels and after ReLU and max-pooling results in 64 feature maps of dimension 16×16. The feature maps are flattened into a vector of size $64 \times 16 \times 16 = 16384$, ABCD and GLCM features are concatenated to this vector which explains why the dimension is actually $1 \times 16384 + 8 = 16392$. The rest of the network is 3 fully connected layers with ReLU activations.

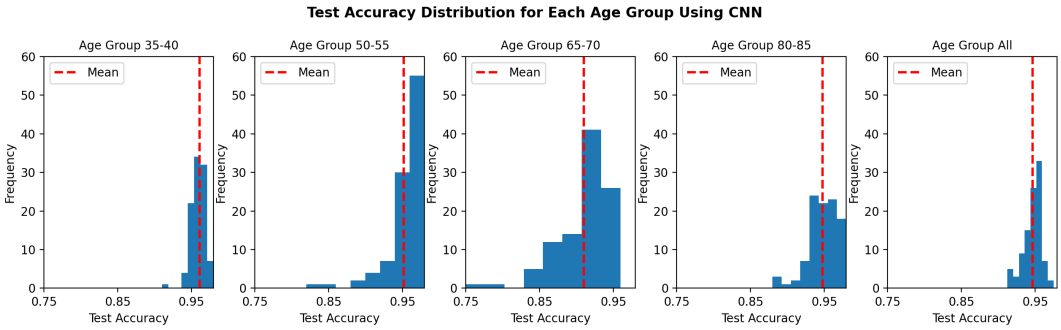

Figure 6: Histograms of Test Accuracy Distribution for Each Age Group Using CNN

Similarly to the XGBoost models, the performance of the CNN models varies across different age groups in terms of both test accuracy and cross-entropy loss. As shown in Figure 6, the mean test accuracy 96.5% for age group 35-40 is slightly higher than 95% accuracy of the overall model. In contrast, with an average accuracy of 92%, the model performance for age 65-70 is less effective than that of the overall model, as presented by the red dashed line.

At the same time, as shown in Figure 7, the mean cross entropy loss for age groups 35-40, 65-70, and 80-85 are around 0.3, 0.3, and 0.35, respectively. They are slightly higher than that of the overall model. In contrast, the model performance for age 50-55 matches that of the general model, as indicated by the red dashed line.

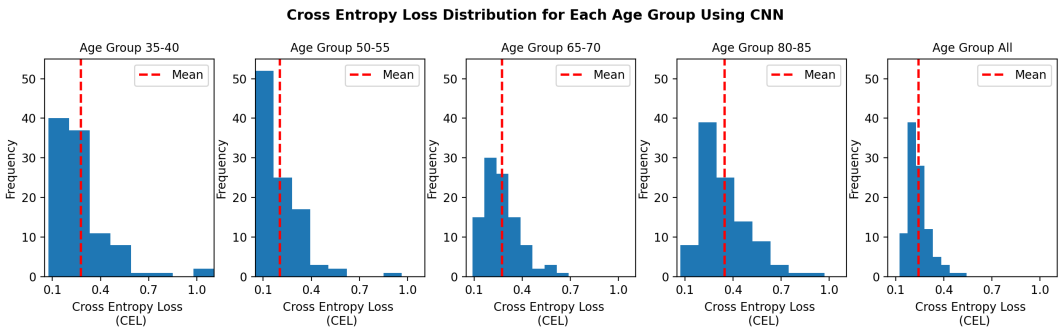

Figure 7: Histograms of Cross Entropy Loss Distribution for Each Age Group Using CNN

Figure 8 provides a side-by-side comparison of distributions between different models. The boxplot of test accuracy on the left clearly demonstrates significant variation among the age groups. In general, all distributions are right-skewed with long left tails. For example, for the overall model, the upper adjacent value is around 96%, the lower adjacent value is approximately 92.5%, and the median is 94%. However, for the age group 35–40, the lower adjacent value is 94%, the upper adjacent value is 98%, and the median is 96%. This indicates that the model for the 35–40 age group explores specific features from its training data, leading to improved performance, while the overall model focuses on broader patterns, resulting in comparatively lower test accuracy.

Next, the boxplot on the right highlights a substantial variability in cross entropy loss among different age groups. In addition to the patterns observed in Figure 8, it is evident that the distribution of each age-specific model is broader than that of the overall model. This may be due to the fact that the general model has more training observations, resulting in more stable predictions.

Finally, a formal p-value test is performed to determine whether all models have the same accuracy level. The test statistic is 1390 which follows an F distribution. With a p-value 1.68×10^{-267}, we conclude that the model performs differently between age groups. In other words, the CNN model suggests a potential to improve skin cancer detection tailored to specific age groups.

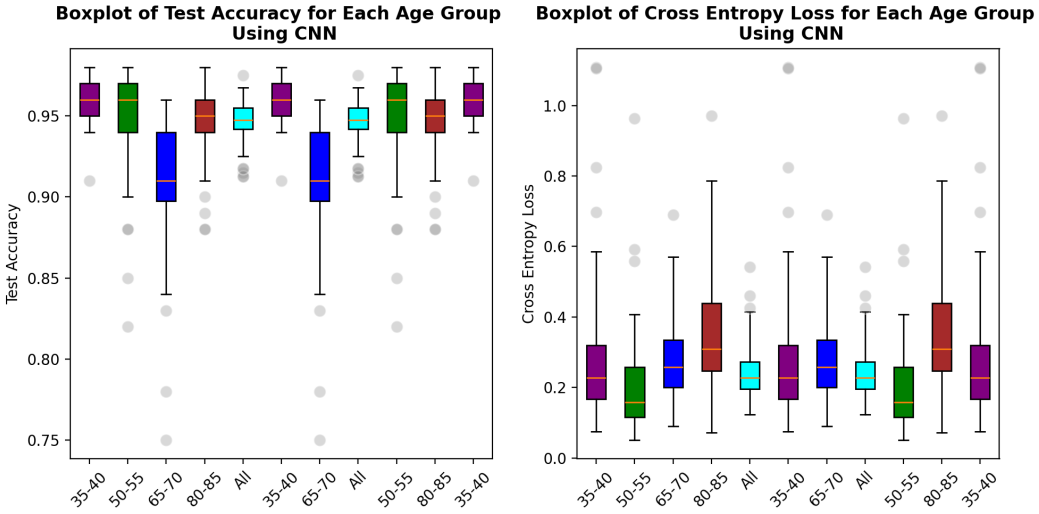

Figure 8: Boxplots of Test Accuracy (Left) and Cross Entropy Loss (Right) Distribution for Each Age Group Using CNN

6 Conclusion and Discussion

In conclusion, both XGBoost and CNN outperform logistic regression. However, logistic regression with LASSO achieved a higher accuracy and lower cross entropy loss than we expected. This may be due to the fact that there exist some critical predictor variables (e.g., pixels in the area of skin lesion), along with several ABCD and GLCM features, that play a significant role in determining whether an image is benign or malignant.

Incorporating age information in separate models improves accuracy for some age groups compared with the overall model's performance. For XGBoost, age groups 50-55 and 80-85 outperformed the overall model, while the age group 35-40 shows no obvious improvement. For CNN, age groups 35-40 and 50-55 outperformed the overall model, while age group 80-85 shows no obvious improvement. However, for both XGBoost and CNN, the 65-70 age group performs worse than the overall model consistently for both XGBoost and CNN. We speculate that this may be caused by the overall model performing better for other age groups which increases the average across groups even though age group 65-70 may actually perform worse than others.

If we compare the performance of XGBoost and CNN, we see that CNN performs better than XGBoost for age groups 35-40 and 80-85, while XGBoost gives higher accuracy for age group 50-55. For the overall model, CNN achieves slightly better accuracy than XGBoost. This agrees with our expectation of model performance between CNN and XGBoost and shows that deep learning does not always dominate over classical models.

7 Future Works

While we obtained good results for all models, there is definitely room for improvement and more in-depth work.

Many of the images have hair that can add noise to ABCD and GCLM features, which consequently affect model parameters. Therefore, we can look into augmenting the data by incorporating segmentation images provided by ISIC and run the two feature extraction methods on these preprocessed images instead. Hyperparameter tunings are done manually in this project, and we can definitely use built-in tools like GridSeachCV from scikit-learn or more advanced heuristic-based search from optimization frameworks like Optuna.

Currently, we are only classifying images based on whether they are benign or malignant. We can go one step further and try to detect whether a malignant skin cancer is melanoma, which is the most common and deadliest among several skin cancers present in the ISIC gallery.

References

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., and Yap, M. H. (2022). Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75:102305.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Codella, N. C. F., Gutman, D. A., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N. K., Kittler, H., and Halpern, A. (2017). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006.

ISIC Archive. Isic archive challenges. `https://challenge.isic-archive.com/challenges/`. Accessed: 2024-10-08.

ISIC Archive. Isic archive gallery. `https://gallery.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D`. Accessed: 2024-10-08.

Kimball, A. B. and Resneck, J. S. J. (2008). The us dermatology workforce: a specialty remains in shortage. *Journal of the American Academy of Dermatology*, 59(5):741–745.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Monika, M. K., Arun Vignesh, N., Usha Kumari, C., Kumar, M., and Lydia, E. L. (2020). Skin cancer detection and classification using machine learning. *Materials Today: Proceedings*, 33:4266–4270. International Conference on Nanotechnology: Ideas, Innovation and Industries.

Skin Cancer Foundation (2023). Skin cancer facts. Accessed: 2024-10-08.

V., B. S., Unnikrishnan, A., and Balakrishnan, K. (2012). Gray level co-occurrence matrices: Generalisation and some new features. *CoRR*, abs/1205.4831.