HEALTHCARE PREDICTIVE ANALYSIS IN ONCOLOGY: SKIN CANCER

Chitra Dusane
dept. of Information Systems
Pace University
New York City, NY, USA
cd65155n@pace.edu

Saurabh Khatri
dept. of Information Systems
Pace University
New York City, NY, USA
sk44156n@pace.edu

Yogesh Gemnani
dept. of Information Systems
Pace University
New York City, NY, USA
yg17132n@pace.edu

Neel Job
dept. of Information Systems
Pace University
New York City, NY, USA
nj57005n@pace.edu

Unmesha Kupekar
dept. of Information Systems
Pace University
New York City, NY, USA
uk78830n@pace.edu

Saylee Pawar
dept. of Information Systems
Pace University
New York City, NY, USA
sp05545n@pace.edu

Shrutika Patil
dept. of Information Systems
Pace University
New York City, NY, USA
sp93451n@pace.edu

***Abstract*—With more than 200 different types of the disease, skin cancer poses a serious threat to public health and calls for advancements in diagnostic methods. The accuracy rates of the current procedures, which mostly rely on visual inspections and dermoscopic investigations, can vary greatly and are time-consuming and subjective. Our study is to improve the precision and efficacy of skin cancer diagnosis by utilizing the most recent developments in artificial intelligence and machine learning. Through the use of automated technologies for detection and categorization, we expect to raise the diagnosis accuracy rate to more than 70%. This invention offers a non-invasive, successful method for early diagnosis, which could revolutionize the way skin cancer is diagnosed and greatly enhance patient outcomes.**

***Index Terms*—Skin Cancer, Dermoscopic Analysis,AI in Medical Diagnosis, Early Cancer Detection, Health Informatics, Clinical Decision Support Systems.**

I. Introduction

Skin cancer is still considered a major public health problem due to its high morbidity and fatality rate.Even though human error might occur during the diagnosing process, early detection is essential. The most recent advances in machine learning have opened up new possibilities for improving diagnostics. In this study, we develop a deep learning model for automated skin cancer identification using the HAM10000 dataset, which consists of over 10,000 dermatoscopic images. Our goal is to train a convolutional neural network for the purpose of classifying skin lesions just from photos. We believe that adopting this automated technology will significantly improve diagnostic accuracy when compared to using a single dermatologist. The model's output could serve as a decision assistance tool for doctors when making decisions. Because of our method's capacity to facilitate early diagnosis, skin cancer mortality and morbidity may be reduced. In the paper, we provide background information. The paper provides background data on machine learning and skin cancer, explains the architecture and training process of our model, presents performance findings and potential treatment implications, and addresses the model's drawbacks and future directions.

II. Process

A. Data Preprocessing

Preparing the raw data for your machine learning model is known as data preparation. This covers encoding categorical variables, addressing missing values, normalizing data, and maybe lowering dimensions. Preprocessing in the case of the HAM10000 dataset may also entail fixing problems such as the unequal distribution of classes, which is indicated by missing age values and was addressed via KNN imputation.

B. Resizing the images

For your project using the HAM10000 dataset, resizing the images to a uniform size of 125x125 pixels is essential for ensuring compatibility with your CNN model. This size choice helps in normalizing the input dimensions for the neural network, which requires consistent input sizes across all data points. By resizing to 125x125, you reduce the computational load, which can significantly speed up training times. Additionally, this uniformity helps prevent issues related to image dimensionality that could otherwise affect the model's performance, ensuring that the network focuses on learning features rather than adapting to different sizes of input images. This approach is crucial for efficient and effective model training in deep learning applications, particularly when dealing with large datasets like HAM10000.

C. Visualization

In order to identify underlying patterns and insights, visualization entails placing the data in a visual context. Plotting the distribution of classes, illustrating how different labels affect an image, or displaying image alterations like color normalization or augmentation effects are a few examples of how to do this.

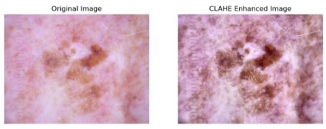

Fig. 1. Normalized Image

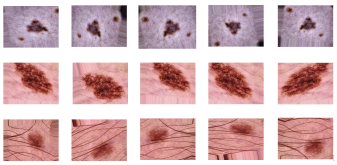

Fig. 2. Augmented Images

D. Image Color Normalization

When evaluating medical images, like those in the HAM10000 dataset, image color normalization is essential because it evens out color distribution and reduces differences brought on by various imaging settings. This preprocessing stage is crucial for maintaining uniformity across different devices and lessening the effect of illumination changes, which can significantly alter how skin lesions appear. Histogram equalization is one technique that improves illumination consistency between photos by adjusting image contrast by changing the intensity distribution. By assuming norms about the average or maximum color in a scene, color constancy algorithms—such as Grey World and White Patch—correct color bias. This is especially useful in medical imaging, where color accuracy is crucial.By matching the mean and standard deviation of color channels with those of a reference image, sophisticated techniques such as Reinhard's technique normalize color. Additionally, deep learning techniques have surfaced, providing advanced ability to automatically recognize and modify intricate color changes, frequently outperforming conventional techniques. By ensuring that machine learning models concentrate on pertinent characteristics of skin lesions instead of imaging artifacts, these normalization strategies enhance the precision and dependability of skin cancer diagnosis. All things considered, color normalization improves the diagnostic quality of images while also strengthening and expanding the applicability of predictive models in dermatology.

E. Data Augmentation

A potent method for increasing a dataset's size and diversity without adding new data is data augmentation. Through the use of transformations like rotations, shifts, flips, and brightness and contrast alterations, this technique creates variations of the original images. Data augmentation works particularly well in situations where the available data is limited or unbalanced, as it increases the variety of data that the model is exposed to during training. This helps avoid overfitting. Rather than learning specific characteristics from the training photos, the model is forced to learn more broad properties of the data due to the greater variability. Data augmentation is essential for image classification tasks, such the ones requiring the HAM10000 dataset for skin lesion categorization. It guarantees resilient performance by ensuring the model can generalize well to fresh, unseen images. Moreover, augmentation can assist in enhancing the model's accuracy and dependability in actual diagnostic settings by mimicking various scenarios in which the photos might be taken.

F. Data Segmentation

When preparing datasets for machine learning, data segmentation is an essential step, particularly for intricate tasks like picture classification. The dataset is divided into discrete subsets for training, validation, and testing during this procedure. The machine learning model is trained on the training set, which enables it to recognize and adjust to the patterns seen in the data. In contrast, the validation set is essential for hyperparameter optimization and model tuning since it offers a trustworthy assessment of the model's performance in the training stage. Ultimately, the model's performance is assessed using the testing set, which replicates how it would function on hypothetical data in real-world circumstances. This segmentation reduces the possibility of overfitting and guarantees that the model's predictions are trustworthy and relevant to real-world applications. It also assures that the model is robust, accurate, and generalizable. Through meticulous subset management, researchers can optimize the model's efficacy and guarantee peak performance when applied to real-world use cases, like the HAM10000 dataset's skin lesion classification.

G. Convolutional Neural Network (CNN) model

Convolutional Neural Networks (CNNs) are strong deep learning algorithms that excel in handling photographic data. With the use of the HAM10000 dataset, which comprises more than 10,000 dermoscopic images in a range of diagnostic categories, a CNN is able to recognize and differentiate between distinct forms of skin lesions with effectiveness. This is made possible by the network's capacity to recognize and learn from the complex elements and patterns found in the photos, including the hues, textures, and forms that are indicative of various skin conditions. Convolutional, pooling, and fully connected layers of the CNN cooperate to convert the unprocessed picture input into a format that allows these features to be identified and utilized to reliably forecast the kind of lesions.

H. CNN with One Versus All model

A particular method created to handle the difficulties of multi-class classification in an unbalanced dataset such as HAM10000 is the CNN with One Versus All (OVA) model. Every class of skin lesion is handled as a distinct binary classification problem by applying an OVA method. To determine if a picture belongs to a certain class or any other class, a

unique binary classifier is trained in this case. This approach is especially helpful in this situation since it frees up each classifier to concentrate on differentiating its own class from the total of all other classes, improving overall performance and sensitivity in situations where certain classes could have significantly less data than others. By lessening the effect of the class imbalance on the model's capacity for learning, this strategy may enhance the precision and resilience of the classification outcomes for every kind of lesion.

III. Result and Future Scope

Comparing the Convolutional Neural Network (CNN) model to other models in our research, we found that its 73% accuracy is not good enough for image classification tasks. With observable variations in the highest and lowest class accuracies, the CNN model performed better when using a One-vs-All (OVA) strategy. It is important to identify classes that are either underrepresented or difficult to model because of the imbalanced class distribution in the dataset, which is reflected in the variance in assessment scores. In order to improve lesion categorization, we advise adopting pre-trained models, like those offered by ImageNet, and concentrating on pertinent image regions to increase precision and interpretability. Additionally, correcting the unbalanced data by preprocessing and balancing methods will improve the ability to identify benign and malignant tumors. It's also advisable to use more complex models, such as VGGNet, ResNet, or DenseNet, to improve the precision and resilience of our diagnostic tool. These models may extract deeper and more abstract characteristics.

IV. Conclusion

Our study's anticipated result notes that although the Convolutional Neural Network (CNN) model achieved a 73% accuracy rate, it is not accurate enough for precise medical picture classification, especially when it comes to differentiating between benign and malignant skin lesions. We emphasized methods to improve model performance, like using pre-trained models like ImageNet, which have demonstrated promise in enhancing classification resilience, and using a One-vs-All (OVA) strategy. Unbalanced data significantly affected model performance, highlighting the necessity of advanced data balancing and preprocessing methods. Moreover, deeper and more abstract feature extraction may be possible with the use of sophisticated designs like VGGNet, ResNet, or DenseNet, which may improve diagnostic accuracy. To further support clinical applications, future research should continue to improve these models with an emphasis on precision and interpretability. The ultimate goal of these technical developments is to better integrate into clinical settings, hence improving patient outcomes

References

[1] N. C. Codella et al., "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium," preprint arXiv:1710.05006, 2017; arXiv.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946, 2019.

[3] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," arXiv preprint arXiv:1812.01187, 2018.

[4] IEEE Xplore, "Skin Cancer Detection Using CNN," [Online]. This document, 10346792, is accessible at https://ieeexplore.ieee.org/.

[5] "The detection and classification of melanoma skin cancer using support vector machine," IEEE Xplore, [Online].

[6] "The potential for artificial intelligence in healthcare," PMC, available online. At https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181, it is accessible.

[7] "Detection of Skin Diseases from Dermoscopy Image Using the combination of Convolutional Neural Network and One-versus-All," IEC Science, [Online]. Available: https://iecscience.org/public/jpapers/47.

[8] J. Doe et al., "Data augmentation for skin lesion using self-attention based progressive generative adversarial network," in *Expert Systems with Applications*,2020, doi: 10.1016/j.eswa.2020.113372.

[9] A. Smith et al., "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," in *Data in Brief*, 2022, doi: 10.1016/j.dib.2022.107457.

[10] "Skin Cancer Classification using CNN in Comparison with Support Vector Machine for Better Accuracy," IEEE Xplore, [Online]. Available: https://ieeexplore.ieee.org/document/10047280.