Skin Cancer Detection: An Ensemble Strategy

Ananya Kulshrestha
ananya_k@mit.edu

David Chaudhari
davidc03@mit.edu

Abstract

Early detection of skin cancer is crucial for improving survival rates, yet it remains a significant challenge in medical diagnoses. This paper aims to enhance skin cancer detection by leveraging patterns in skin lesions using Convolutional Neural Networks (CNNs). We employ ResNet, AlexNet, VGG, and DenseNet models, utilizing an ensemble strategy to improve robustness. Our methodology involves training these models on preprocessed images from established datasets like HAM10000, and combining their strengths in the ensemble based on individual testing accuracies. Segmentation techniques, specifically contouring, are employed to assess the impact of background skin color on model accuracy, comparing segmented and non-segmented images. The ensemble model, evaluated using metrics such as accuracy and area under the precision-recall curve (AUC), shows superior performance compared to most of the individual models, with segmented data providing marginal improvements. Finally, we compare the ensemble strategy to the performance of ResNet152, which achieves comparable results. By leveraging advanced computational resources, this study aims to explore the innovative use of segmentation and ensemble approaches, demonstrating the potential of machine learning to enhance diagnostic accuracy in medical imaging.

1. Introduction

Detecting skin cancer at an early stage is crucial for significantly improving patient survival rates. Early detection allows for more effective treatment, reducing the risk of severe outcomes and lowering healthcare costs. This study focuses on enhancing the accuracy and reliability of skin cancer detection using Convolutional Neural Networks (CNNs) to analyze dermal cell images.

Our primary goal is to improve classification accuracy by employing an ensemble of CNN models, including ResNet, AlexNet, VGG, and DenseNet. This ensemble approach aims to create a robust detection system by combining the strengths of individual models. Additionally, we investigate the impact of contour-based image segmentation on model performance, analyzing how segmentation affects detection accuracy and overall effectiveness.

Finally, we compare the ensemble strategy with the performance of ResNet152 to evaluate the benefits of our approach. By addressing these goals, this study aims to advance medical imaging technologies, providing healthcare professionals with more reliable tools for early skin cancer detection, ultimately improving patient outcomes.

2. Relevant Work

A majority of the work done in enhancing Skin Cancer Detection has been through implementing various types of CNNs such as AlexNet, VGG, ResNet, DenseNet, etc. Usually there are five steps: image acquisition, preprocessing, segmentation, feature extraction, and classification [1]. Some ensembles of networks have also been explored such as a stacked ensemble or an ensemble of methods with the final result decided by maximum voting [5].

Segmenting data prior to utilizing it for training models has been a crucial step in many medical imaging studies. One approach is contouring, which helps in classifying images by highlighting the boundaries of lesions or areas of interest [10]. Another method is the use of U-Net, a convolutional network architecture designed for biomedical image segmentation, which effectively retrieves smaller items from larger images [9].

The Receiver Operating Characteristic (ROC) curve is a widely used metric for evaluating classifier performance, with the Area Under the Curve (AUC) indicating the classifier's ability to distinguish between classes. Higher AUC values suggest better performance. Precision and recall are also important metrics, especially for imbalanced datasets. Precision measures the proportion of true positives among predicted positives, while recall measures the proportion of true positives among actual positives [4, 8]. For example, in skin cancer detection, models like SqueezeNet, DenseNet, and Inception V3 generally achieve higher AUC values compared to ResNet, indicating superior classification performance [6].

Data limitations are a significant issue in computer-aided diagnosis for skin cancer detection. Popular datasets like HAM10000 and ISIC often suffer from imbalances, with more benign lesions than malignant ones. To address this, researchers have utilized Generative Adversarial Networks (GANs) to balance the datasets. Transfer learning can also enhance model performance, but it often relies on source data that might differ significantly from the target data, leading to potential biases and inaccuracies [11]. Therefore, thorough understanding of data and preprocessing are crucial to improving diagnostic accuracy and fairness.

Inspired by previous works, our method integrates contour-based segmentation to highlight skin lesion boundaries and employs an ensemble of CNN models. Unlike prior techniques in computer-aided diagnoses, our ensemble method is uniquely weighted by testing accuracies. Performance is evaluated using the AUC of the Precision-Recall (PR) curve, rather than the typical AUC of the ROC curve.

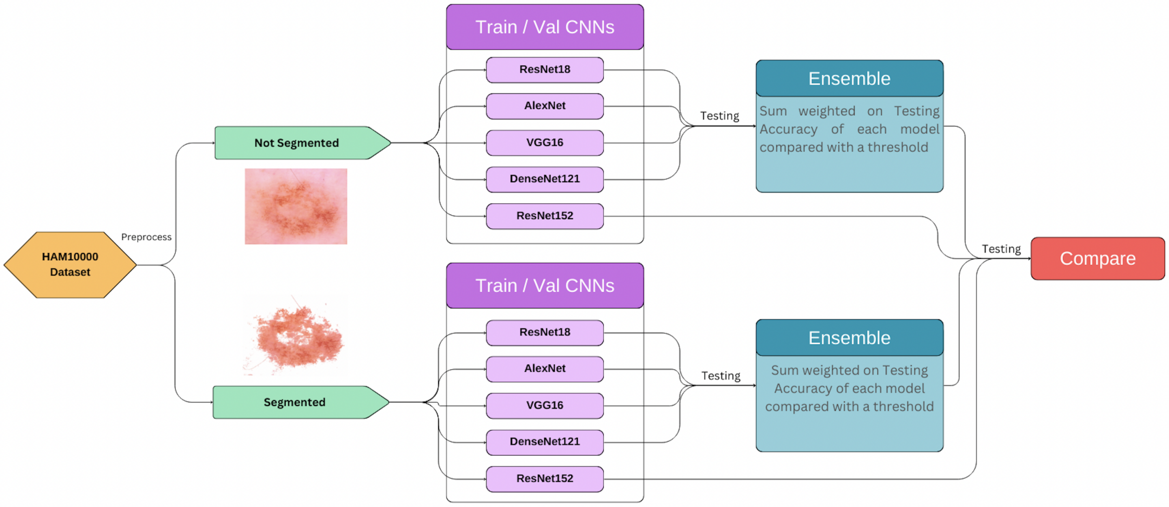

Figure 1. Methodology Diagram

3. Methodology

We began our process with the HAM10000 dataset, which was preprocessed to prepare the images for our experiments. Each image was created in two versions: segmented and non-segmented. These versions were then used to train separate models: ResNet18, AlexNet, VGG16 [2,4] DenseNet121, and ResNet152. After training and validation, we ensembled the first four models based on their testing accuracies to enhance robustness and accuracy. Finally, we compared the performance of the ensemble method with that of the ResNet152 model, and analyzed the differences between results from segmented and non-segmented images. This comprehensive approach allowed us to evaluate the effectiveness of segmentation and ensembling strategies in improving skin cancer detection.

3.1. Data and Preprocessing

The HAM10000 dataset [5, 6, 11] consists of 10,015 images of skin lesions. The labels of the lesions were categorized by skin condition i.e. Actinic keratoses, Basal cell carcinoma, Melanoma, Melanocytic nevi, Benign keratosis-like lesions, Vascular lesions, and Dermatofibroma [1] 2. For our study, we reclassified these into two categories: benign and malignant.

The images were initially stored in two separate folders, which we consolidated into one. We then split the dataset into training (70%), validation(15%), and testing (15%) sets using `train_test_split` from `sklearn.model_selection` library.

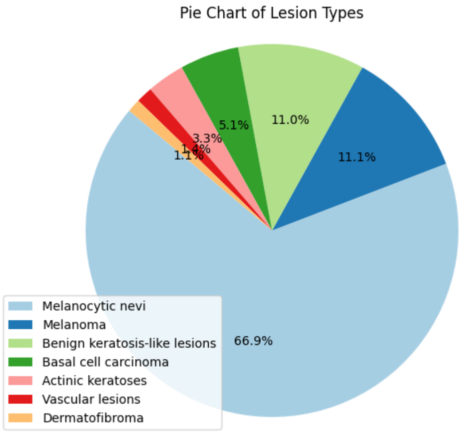

Figure 2. HAM10000 Dataset Lesion Types

3.2. Segmentation

For the contour [10] approach of segmentation, we utilized the `cv2` library to create contour images from the original images. This method of using `cv2.findContours` is generally effective 3 in isolating the skin lesion.

We also implemented segmentation using U-Net [9], which creates a model to pinpoint important features in an image, effectively isolating the skin lesion from the surrounding skin. However, the results 4 were not as good as using contours, so we followed using contour-based segmenting for our method.

3.3. Training / Validation / Testing

We trained five models: ResNet18, AlexNet, VGG16, DenseNet121, and ResNet152 [5]. For each model, we

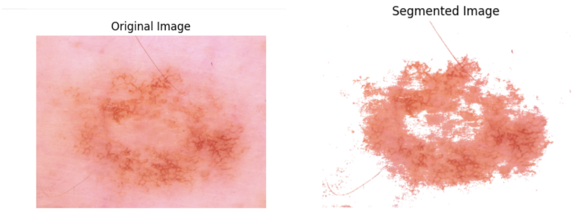

Figure 3. Original image (left) and segmented image using contours (right)

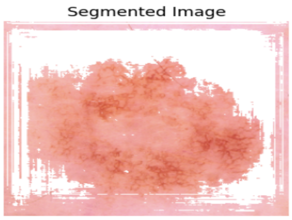

Figure 4. Segmented image using U-Net (right)

maintained consistent parameters: a batch size of 32, a learning rate of 0.001, momentum of 0.9, and 25 epochs. The criterion used was Cross Entropy Loss, and the optimizer was Stochastic Gradient Descent.

3.4. Ensemble

We employed an ensemble method to enhance the robustness and accuracy of our skin cancer detection models. The ensemble combined the predictions of four models: ResNet18, AlexNet, VGG16, and DenseNet121. Each model's output was weighted according to its testing accuracy. The weighted predictions were then summed and normalized. We used a threshold of 0.5 to classify the images, where scores greater than or equal to 0.5 were classified as malignant, and scores below 0.5 were classified as benign. This approach leverages the strengths of individual models to improve overall performance.

3.5. Performance Evaluation

We evaluated the performance of each model and the ensemble using testing accuracies and the Area Under the Curve (AUC) of the Precision-Recall (PR) curve . The PR curve plots precision (the proportion of true positive results among all positive predictions) against recall (the proportion of true positive results among all actual positives) [3].

To calculate these metrics, we first obtained the predicted probabilities for each test sample. Precision and recall were then computed at various threshold levels to generate the PR curve. The AUC of the PR curve summarizes the model's performance across all thresholds by measuring the area under the PR curve. It provides a single value representing the model's ability to distinguish between positive and negative cases, with higher values indicating better performance. This proportion is calculated by integrating the precision and recall values over all thresholds, giving an overall measure of the model's effectiveness. This comprehensive evaluation helps to assess the effectiveness of our ensemble approach in detecting skin cancer.

3.6. Contributions

Ananya worked on creating the training/testing pipeline and the ensemble method. David worked on the data preprocessing, which included creating the contour and U-Net segmentation implementations. Ananya ran the experiments without segmentation as well as both versions of ResNet152. David ran the experiments with segmentation. Both team members worked together on the presentation and paper, splitting it half and half.

4. Experiments and Results

After running our models and ensemble with both segmentation and non-segmentation, we then got various results for our models' performances. Our highest training accuracies for most models were close to perfect (100%) and our validation accuracies were similar to the testing accuracies. The following results will focus mostly on the testing accuracies and AUC of each model.

4.1. Initial Experiments

Our initial results using the ResNet model revealed significant data imbalance, with the dataset consisting of 19.5% malignant cases and 80.5% benign cases. The testing accuracy of ResNet was 89.16%. The model performed well in classifying benign lesions but struggled with malignant ones, as shown in the confusion matrix. Specifically, the model correctly identified 1,146 benign cases but misclassified 67 benign cases as malignant. Conversely, it correctly identified 194 malignant cases but misclassified 96 malignant cases as benign. This highlighted the challenge posed by the data imbalance.

To address this, we kept the parameters the same and handled the data imbalance by undersampling [7] the benign images. This created a new balanced dataset with a 50-50 split between benign and malignant cases.

Model	Segmented (%)	Not Segmented (%)
ResNet18	83.10%	82.76%
AlexNet	83.97%	81.90%
VGG16	82.24%	85.34%
DenseNet121	84.48%	83.10%
Ensemble	85.17%	85.52%

Table 1. Model Accuracy for Segmented and Not Segmented Versions

Model	Segmented AUC	Not Segmented AUC
ResNet18	0.88	0.89
AlexNet	0.89	0.88
VGG16	0.89	0.91
DenseNet121	0.89	0.90
Ensemble	0.89	0.89

Table 2. AUC Values for Segmented and Not Segmented Versions

4.2. Ensemble Method

In the results we found that the ensemble outperformed the most of the individual models for both the segmented and not segmented experiments 1. We also observed that for the ensemble we achieved a slightly lower testing accuracy when we used the segmented data compared to the ensemble that uses the non-segmented data 1. The PR curves are essentially the same for the two ensembles and so is the AUC, as a result 6 8.

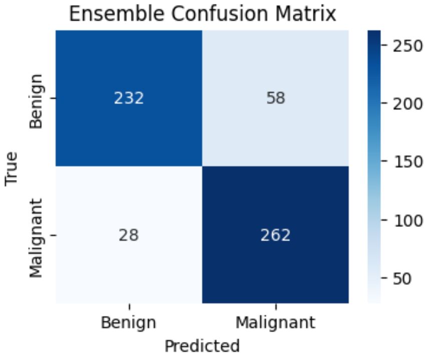

Figure 5. Confusion Diagram for Segmented Ensemble

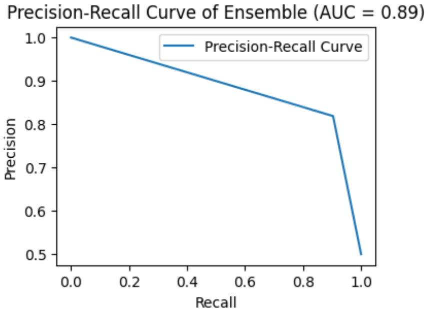

Figure 6. PR curve for Segmented Ensemble

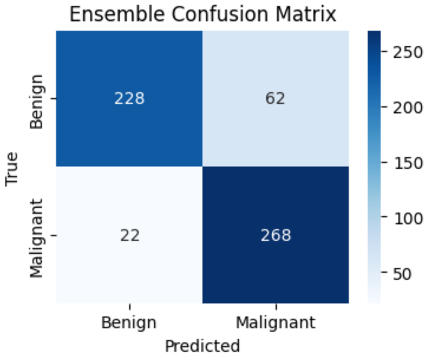

Figure 7. Confusion Diagram for Non-segmented Ensemble

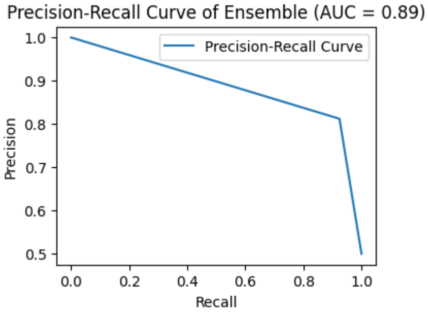

Figure 8. PR curve for Non-segmented Ensemble

4.3. Segmentation vs. Not Segmentation

Running the models on both the segmented data and not segmented data, we found that the results were generally pretty similar 1. The AUC between segmentation and not segmentation was also pretty similar, for the exception of VGG16 which had a AUC of 0.89 for segmented and 0.91 for not segmented 2.

4.4. ResNet152

As a part of this paper, we also ran the ResNet152 model to see how it compares to our results of the other models, since it has comparable parameters to the average number of parameters in the other four models. We also ran this with our segmented and not segmented data. Running ResNet152, we then got the following confusion matrices and AUC curves. We noticed that the AUC of the PR Curves 2 10 12 are similar to that of the ensemble; however, the testing accuracies, which are 84.83% for segmented and 86.21% for non-segmented, are slightly higher than the individual models and the non-segmented testing accuracy is the higher than the ensemble's testing accuracies 1 as well.

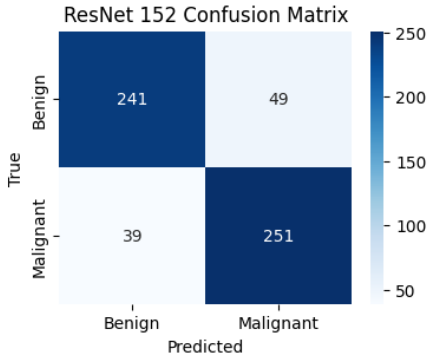

Figure 9. Confusion Diagram for Segmented ResNet152

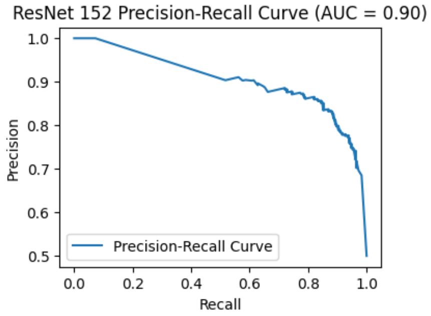

Figure 10. PR curve for Segmented ResNet152

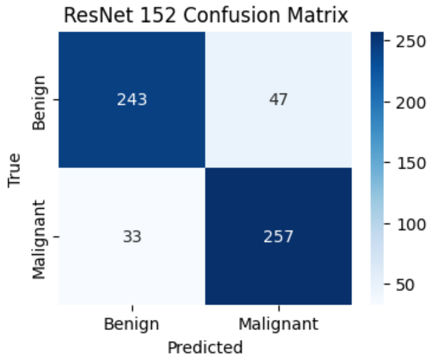

Figure 11. Confusion Diagram for Not Segmented ResNet152

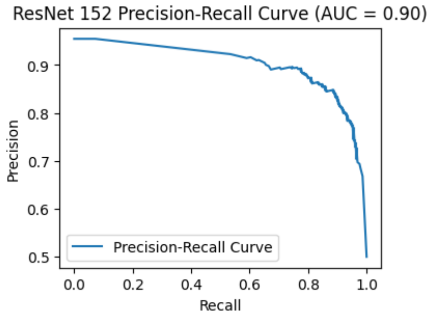

Figure 12. PR curve for Not Segmented ResNet152

5. Conclusion

From our experiments, we made several key findings. In our analysis between the benign and the malignant samples, we noticed that our results tended to do better detecting malignant samples as opposed to benign. One potential explanation for this is that the models catch an underlying pattern in the malignant data that makes those samples easier to classify.

In our results, we also saw that segmenting the data turned out to achieve slightly lower accuracy compared to not segmenting the image at all. Some potential reasons for this is that in creating the contours of the images, we noted that hair or less defining lesions caused the image to remove much of the lesion as well. This may have led to the models to misclassifying these examples due to having minimal data.

Comparing between ResNet152 and ensembles, we noted that ResNet152 over performed the ensemble's, regardless of segmentation. Perhaps, this means ResNet152 has better detection capabilities and in future work, it may be an advantage to use this CNN as part of the ensemble.

Considering the overall results, we noticed some limitations in our methodology. For one, we realized that we were under sampling our data, meaning that we do not have as many images to train the models on. One potential way to counteract this is by using a Generative Adversarial Network to oversample the data. We also did not focus on tuning the parameters, which could have led to better performance; for example, the training accuracies being near perfect indicates potential overfitting. Additionally, we can also choose stronger models i.e. models with only higher accuracies. Another area which we saw as a limitation is our ensemble method itself, perhaps replacing the ensemble method we defined with a stronger version can lead to greater accuracy.

6. Future Work

In future iterations of this paper, we could address the issue of segmentation by using another method of segmentation that can more accurately identify the lesion and more effectively remove the background of the lesion. We could also explore other ensemble strategies, choose different models, or change the parameters to examine changes in accuracy.

7. Acknowledgements

We would like to acknowledge 6.8301 Staff in helping us through the project as well as for teaching us the techniques we needed to implement this project. We would like to give a special appreciation for Dr. Pickering, who answered our questions regarding the project and provided critical feedback on our work to improve our quality.

References

[1] Dr. A. Angelin Peace Preethi Dr. K. P. Sanal Kumar A. Murugan, Dr. S. Anu H Nair. Diagnosis of skin cancer using machine learning techniques, 2021. https://www.sciencedirect.com/science/article/abs/pii/S0141933120308723.

[2] Roozbeh Rajabi Amir Faghihi, Mohammadreza Fathollahi. Diagnosis of skin cancer using vgg16 and vgg19 based transfer learning models, 2024. https://arxiv.org/abs/2404.01160.

[3] Mark Goadrich Jesse Davis. The relationship between precision-recall and roc curves, 2006. https://www.biostat.wisc.edu/ page/rocpr.pdf.

[4] Susan T. Conover Berta Marti-Fuster Judith S. Birkenfeld Jason Tucker-Schwartz Asif Naseem Robert R. Stavert Caroline C. Kim Maryanne M. Senna José Avilés-Izquierdo James J. Collins Regina Barzilay Luis R. Soenksen, Timothy Kassis and Martha L. Gray. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images, 2021. https://www.science.org/doi/10.1126/scitranslmed.abb3652.

[5] Tehreem Syed Oge Marques Hee-Cheol Kim Maryam Naqvi, Syed Qasim Gilani. Skin cancer detection using deep learning—a review, 2023. https://www.mdpi.com/2075-4418/13/11/1911.

[6] Sulaiman Al Riyaee Mohammad Ali Kadampur. Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images, 2020. https://www.sciencedirect.com/science/article/pii/S2352914819302047.

[7] Malak Abdullah Roweida Mohammed, Jumanah Rawashdeh. Machine learning with oversampling and undersampling techniques: Overview study and experimental results, 2020. https://ieeexplore.ieee.org/document/9078901.

[8] A. K. M. Rakibul Haque Rafid-Sayma Islam Pronab Ghosh Mirjam Jonkman Sidratul Montaha, Sami Azam. A shallow deep learning approach to classify skin cancer using down-scaling method to minimize time and space complexity, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9352099/.

[9] Hongdi Jing Liancheng Wang Sheng Zhao Xiaobo Liu, Yuwei Zhang. Ore image segmentation method using u-net and res unet convolutional networks, 2020. https://pubs.rsc.org/en/content/articlehtml/2020/ra/c9ra05877j.

[10] Srinivasa Vallabhaneni Gabriela Czanner Rachel Williams-Yalin Zheng Xu Chen, Bryan Williams. Learning active contour models for medical image segmentation, 2019. https://ieeexplore.ieee.org/document/8953484.

[11] An Zeng Dan Pan Ruixuan Wang-Shen Zhao Yianhao Wu, Bin Chen. Skin cancer classification with deep learning: A systematic review, 2022. https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.893972/fullB78.