Image Classification for Identification of Cancerous Lesion

Joanna Yang , Elad Oz , William Teng , Nitin Hari Gomatam , Alexander Chen

281 Final Project | Summer 2024

1 Abstract

When determining whether or not a skin lesion is cancerous or non cancerous, doctors and dermatologists rely on an initial visual examination of the area before making other decisions on how to proceed[1]. If some of the key variables are concerning, doctors usually perform a biopsy for a more detailed examination[2]. Dermatologists assess several critical variables when evaluating skin lesions. These include asymmetry, where one half of the lesion differs from the other; irregular or poorly defined borders; color variation; size, particularly when the diameter exceeds 6 millimeters; and the degree of protrusion above the skin surface[3]. We extracted features to proxy the variables using HSV, Gabor, Linear Binary Patterns, measuring distance from centroid, reflection score, and features returned from neural network models like ResNet50. For classification we used several methods where random forest and logistic regression returned the best accuracy scores of 75.8% and 75.7% respectively.

2 Introduction

The accurate identification of lesions as cancerous or noncancerous in modern healthcare will influence treatment efficacy and help plan necessary treatment for patients. Early and precise diagnosis help plan timely treatment, particularly critical in conditions like melanoma where early detection improves prognosis, whereas misdiagnoses, including false negatives, can delay necessary interventions.

This project focuses on classifying skin lesions between cancerous and noncancerous classifications by extracting features using techniques such as HSV, Gabor, Linear Binary Patterns, measuring distance from centroid, reflection score and ResNet-50 convolutional neural network which are modeled through Random Forest and Logistic Regression. Authoritative bodies such as the World Health Organization, American Cancer Society, National Cancer Institute, and International Agency for Research on Cancer play pivotal roles in standardizing diagnostic criteria and using the HAM10000 image dataset, we intend to contribute to potential new ways to classify lesions to benefit patients and diagnostic healthcare professionals.

[1] *(Halpern, 2021)*
[2] (*Tests For Melanoma Skin Cancer | Melanoma Diagnosis*, *2023*)
[3] *(How Do I Identify a Suspicious Lesion?, 2020)*

3 Data

Our project focuses on classifying skin lesions from the HAM10000 dataset[4], which consists of 10,015 dermatoscopic images of seven different types of pigmented lesions. Each image has dimensions of 450x600 pixels per color channel and a corresponding segmentation mask for the lesion. Over 50% of the lesions are confirmed through histopathology, while the ground truth for the remaining cases is established through follow-up examinations, expert consensus, or in-vivo confocal microscopy.

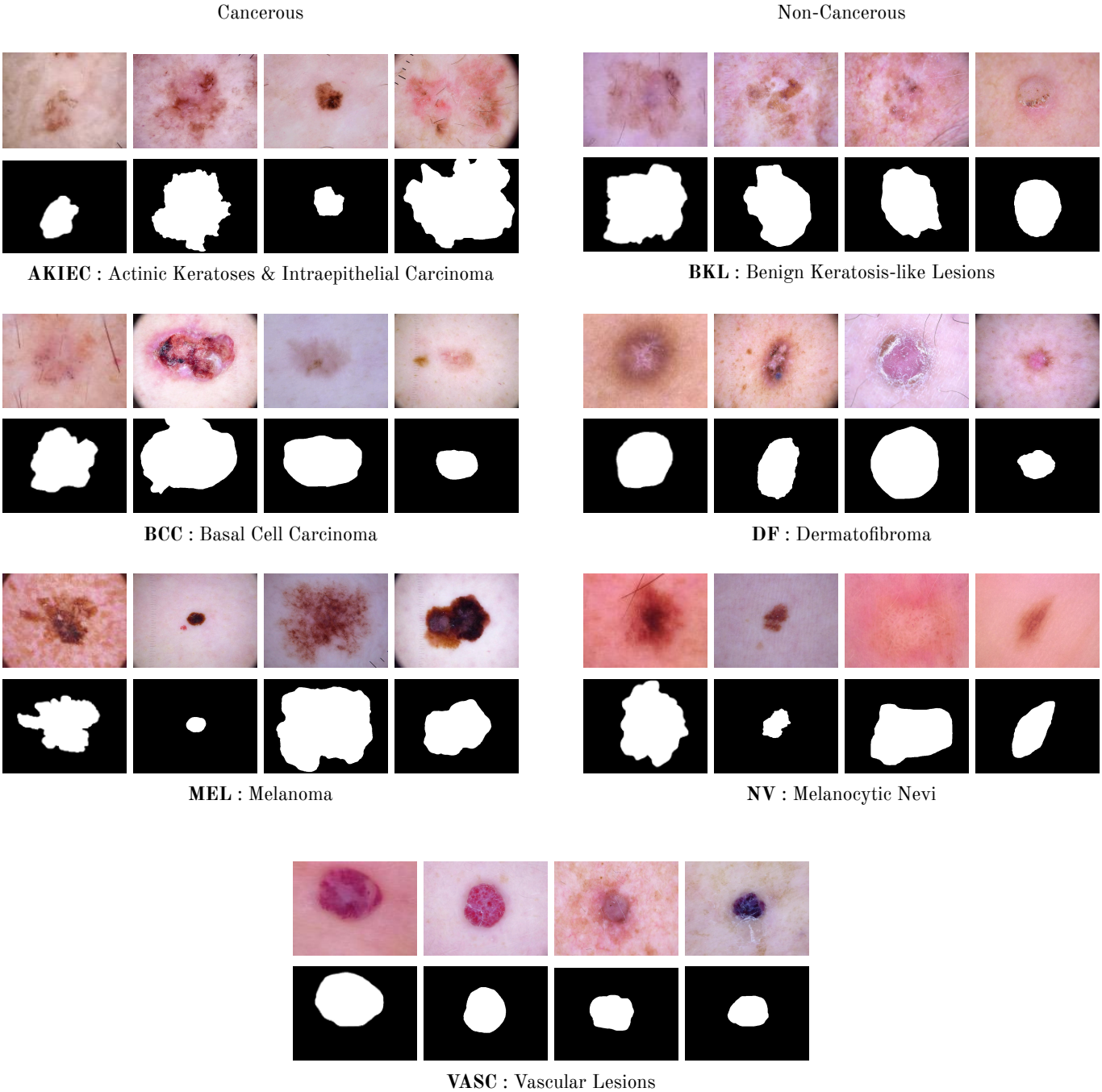


Figure 1. Example images and segmentations from each skin lesion classification.

[4] *(Philipp Tschandl, 2018)*

2 Feature Extraction

The HAM10000 dataset comes with metadata that contains patients' age at time of the image being taken, their sex and localisation, which is where the lesion is on their body. To gather the additional lesion traits intended to be incorporated into our model, we require extraction techniques applied to the images themselves, which will be outlined in the following section.

2.1 Preprocessing

The following section outlines the processing done prior to the feature extraction process.

2.1.1 Hair Removal

Upon reviewing the images, we noticed that a number of the images had hair covering the lesion or ruler markers to measure the lesion. When initially attempting to extract features, such as edges of the lesions, the feature extraction algorithms would pick up characteristics of the hair or ruler markers. Since we wanted to focus on information about the lesion itself, we used an algorithm called Dullrazor[5] to remove the hair.

The algorithm converts the image to grayscale, uses BlackHat to enhance dark objects of interest, blur the output image, and then apply binary thresholding to obtain a mask of the hair. This mask is then used to remove the hair pixels, which is then replaced with image inpainting.

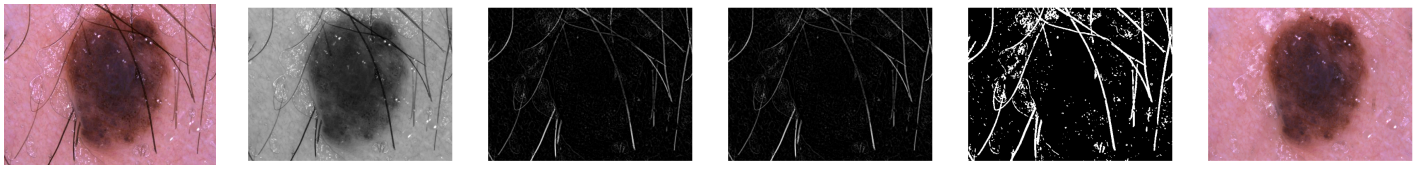


Figure 2. Example of the Dullrazor algorithm on one image.

2.1.2 Property Identification

Due to a wide variation in lesions, despite the lesion itself generally being somewhat in the main focus of the image, in order to apply the metric for features, each property in the masked image is calculated for their respective size and the largest property is determined to be the main lesion to calculate metrics from.

[5] *(Lee et al., 1997)*

2.2 Symmetry

Lesion asymmetry can be an indicator of the lesion's characteristics and potential malignancy. Typically, benign lesions, like nevi, exhibit symmetry, whereas malignant lesions, like melanomas, often present asymmetry[6].

The following steps were taken to calculate symmetry scores on a 0-1 scale for each lesion:

1. Find contours of lesion and extract from the largest lesion
2. Compute bounding box around largest contour and extract lesion region
3. Calculate symmetry scores across the x-axis, y-axis, main diagonal, and anti diagonal by reflecting lesion and compute percentage overlap on a 0-1 scale
4. Average symmetry scores to compute combined symmetry score on a 0-1 scale

The x-axis and y-axis symmetry scores are calculated by splitting the lesion region into top/bottom and right/left halves and flipping each pixel at (i,j) on one half to compare it with the respective pixel on the other half. To calculate the main diagonal symmetry score, for each pixel at (i,j) in the lesion region, it is reflected across the main diagonal to (j,i). To calculate the anti diagonal symmetry score, for each pixel at (i,j) in the lesion region, it is reflected across the anti diagonal to (h—j—1,w—i—1), where h and w are the height and width of the region.

2.2.2 Eccentricity

Cancerous lesions are known to be less circular[7] and to quantify this, eccentricity is a measure used in image processing to quantify how much a shape deviates from being circular. Eccentricity (e) quantitatively assesses the degree to which a shape diverges from a perfect circle, expressed as a value ranging from 0 to 1. A value of e approaching 0 indicates a shape closely matching a circle, whereas values nearing 1 suggest a significant deviation from circularity.

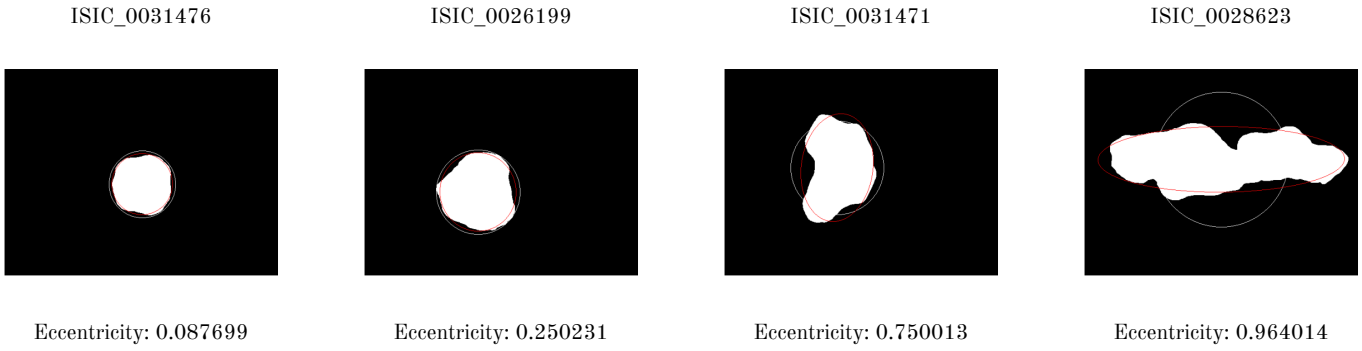


Figure 3. Examples of eccentricity

The eccentricity (e) of an ellipse is calculated using the formula:

$$e = \text{sqrt}(1 - b^2 / a^2)$$

where b is associated with the y-factor for the ellipse and a is related to the x-factor of the ellipse

[6] *(Ali, 2020)*
[7] *(Ali, 2020)*

To determine the significance of eccentricity, the below breakdown of the distribution by each classification can be seen by the distribution. AKIEC, BCC, and MEL are cancerous, whereas BKL, DF, NV, and VASC are not.

Cancerous Lesions

Non-Cancerous Lesions

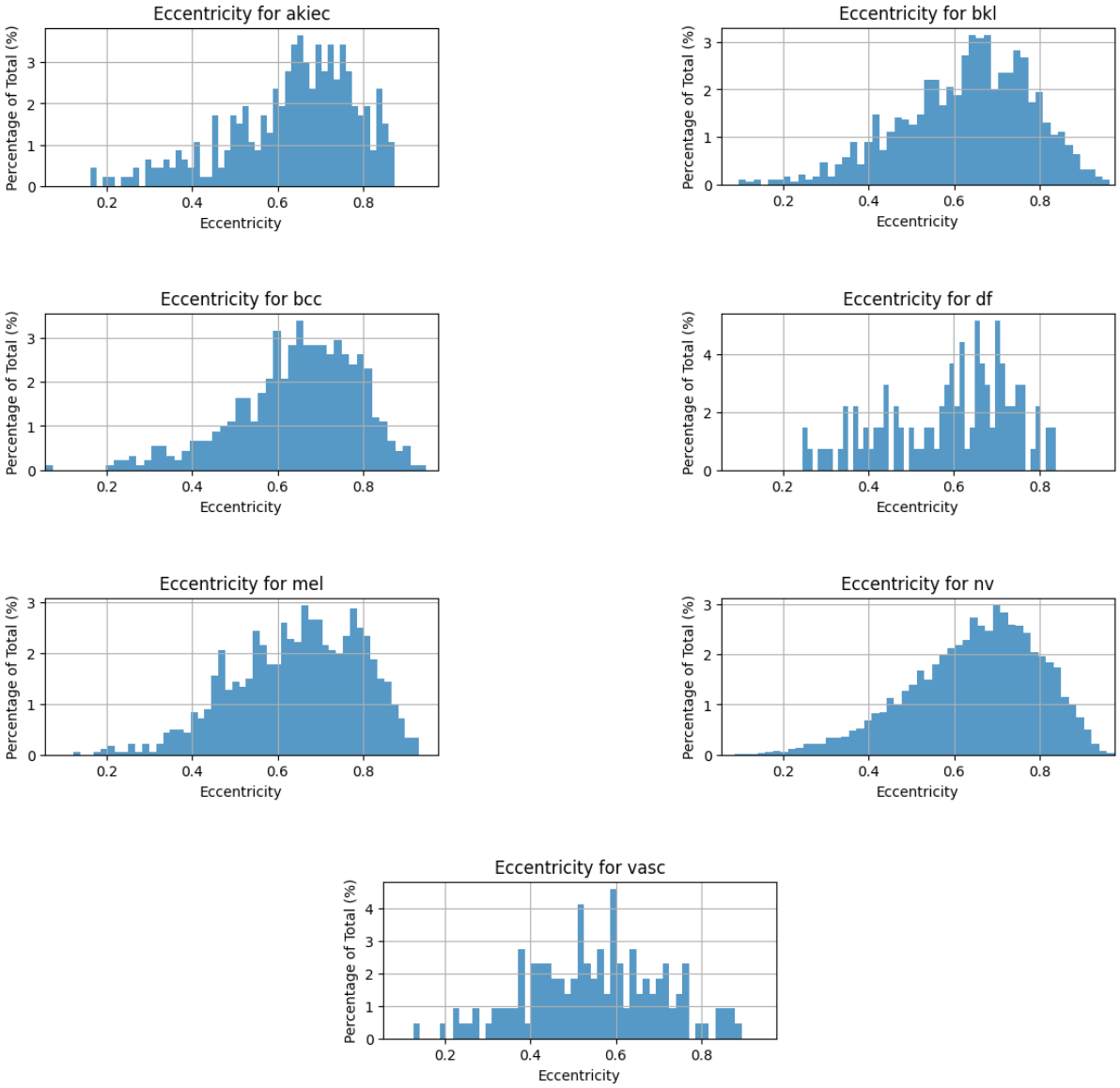


Figure 4. Distribution of eccentricity across classifications

2.2.3 Color

Due to color variation being an indicator of skin lesion malignancy and type[8], we decided to characterize the color variances within the lesion itself and also compared to the surrounding skin.

Histogram of Color

The initial approach taken was to analyze the histogram of color along the axis of hue, saturation, and value to distinguish between skin lesion types. Hue represents the color itself across the red, blue, and green spectrum; saturation reflects the color's intensity or dullness; and value quantifies the color's lightness or darkness.

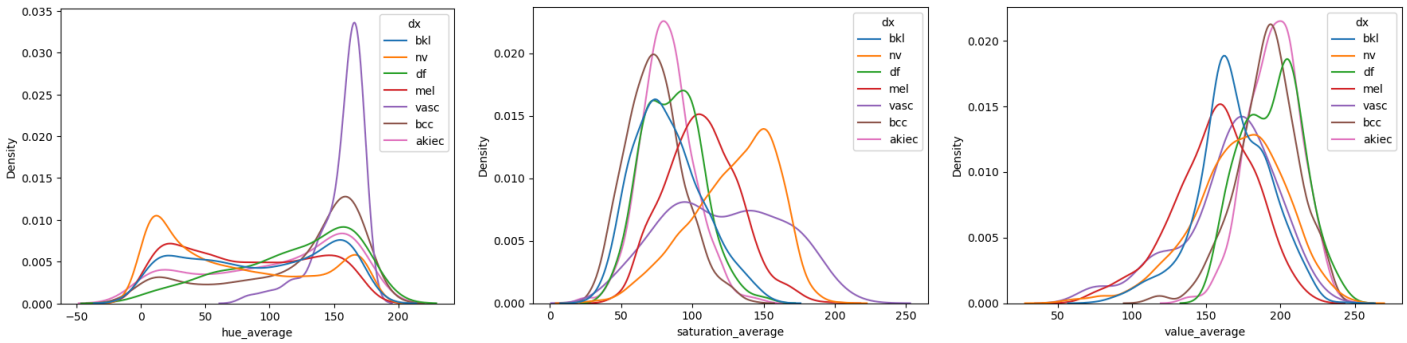


Figure 5. Distribution of HSV averages across groups normalized independently

While most of the distributions appear to be similar, there are a few outliers that distinguish one lesion type from another. For example, in hue, VASC lesions have a peak distinct from other groups and the saturation averages for NV differs from the other groups.

HSV of Lesion v. Surrounding Skin

Another approach involved calculating the mean and median hue within each lesion and comparing these metrics with those of the surrounding skin. For each image and its corresponding mask, we computed the average and median hue values both within and outside the segmented lesion area. This provides insights into the color differences between the lesion and the surrounding skin.

HSV of Color Variances within Lesion Quadrants

Another approach involved dividing each lesion into quadrants and computing the average and median hue for each quadrant. The goal is to identify the darkest and lightest hue pixels within the lesion to help identify if there was any difference in color within each lesion. This would help identify color-based patterns or characteristics that might be relevant for understanding different types of skin lesions.

Some of the data manipulations that we accounted for was removing the color black from the min/max hue calculation, otherwise for most lesions any detection of the color black would lead to that skewing the data.

---

[8] *(Ali, 2020)*

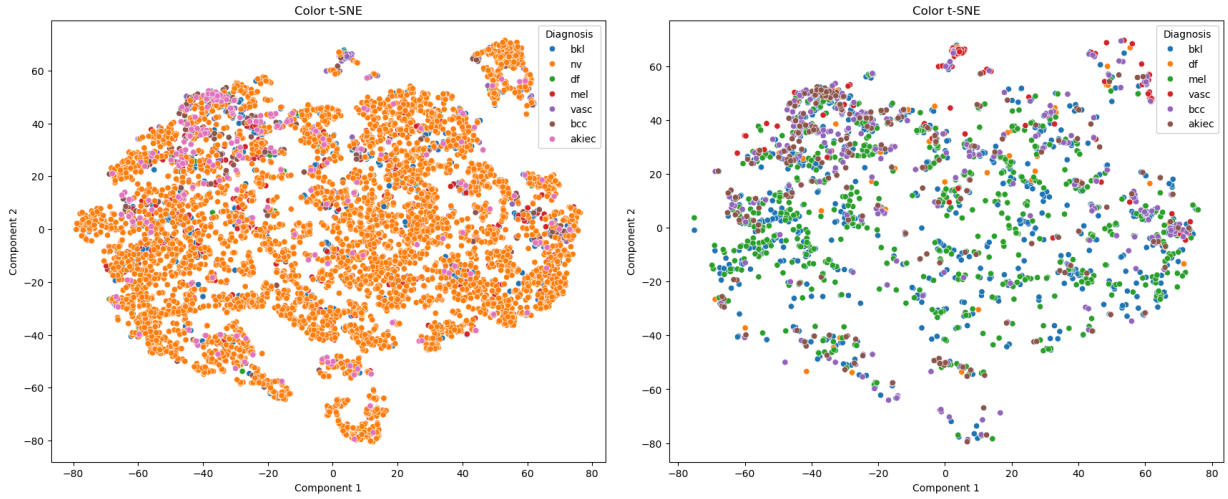


Figure 6. t-SNE of color

Using t-SNE, which is a method of dimensionality reduction to visualize how well a feature is distinguished between classes, it appears that NV covers cases in most classes. This is reasonable because NV is a noncancerous skin lesion that is commonly known as a mole. Having a mole can sometimes warrant further follow-up to diagnose if it is any type of cancerous lesion. However, after removing NV, there appears to be some differences between classes, especially with VASC lesions and AKEIC lesions.

2.2.4 Texture

We make use of classical texture-capturing methods to supply our model with discrimination power beyond that available in color distributions. Texture features are invariant to skin pigmentation and therefore, we hope, contribute a degree of robustness to the model.

The first texture feature we use is a Gabor filter. Gabor filters are a family of parametrized linear filters whose impulse response is defined by a sinusoidal plane wave multiplied by a Gaussian. This constitutes a 2-dimensional band-pass filter, able to identify the significance of frequencies in some predefined bandwidth. Since that bandwidth, along with the direction of measurement are limited in a single filter, a collection of these is usually used, called a filter bank. However, each filter in the bank would lead to another set of features with cardinality matching the number of pixels in the image. To control for the overall number of features in our model, we use a single Gabor filter, whose hyperparameters were optimized visually for discrimination between the classes. By post-processing the output of this convolution (more below), we manage to extract useful information from this single filter.

The second texture feature is the linear binary pattern (LBP). This is an encoding of the region surrounding a pixel obtained by thresholding each neighboring pixel against the pixel in question, and mapping the relationship between the two to 0 or 1. We then collect the resulting values into a binary string and use the numerical value of that binary string as the texture representation of the pixel. This encodes local contrast observed in the vicinity of each pixel. In much the same way as the Gabor filter, hyperparameters were selected visually for discrimination power between images.

The features described above, as with most commonly used texture methods, yield an output whose dimension matches that of the image. To control for the number of features, and ensure features invariance to lesion position and rotation, we aggregate the values produced in the following way. We first leverage the segmentation masks to identify the center and radius of each lesion's largest bounded circle. We then identify the lesion's smallest bounding circle with the same center. We then average the two radii, and use the resulting circle as a set of polar-coordinate axes around the lesion. We bin the radius into 8 equally spaced radii and compute aggregating statistics over each concentric ring. Finally, aggregate over the radial direction, using the following aggregations:

- Standard deviation of ring means
- Kurtosis of ring means
- Skew of ring means
- Interquartile range of entire circle's values
- Radial mean slope: OLS coefficient of each ring's mean regressed on the ring number (size-invariant measure of distance from center)
- Radial standard deviation slope: same as above, but aggregate each ring by the standard deviation of its values.
- Ring number of ring with maximum internal variation (Std)
- Ring number of ring with maximum internal mean
- Lesion region mean relative to outer (non lesion skin) mean

The 18 (9+9) features are then standardized, and then combined with the meta-features and fed into a simple logistic regression classifier for evaluation and features selection. The meta-features alone achieve an accuracy score of 25.6% on the validation set (30% of the dataset). When adding the LBP features, we get 34.9%. The meta and Gabor features combination yields 34.3%, and the combination of meta, Gabor, and LBP features yields 39.2%. A backward stepwise selection algorithm is then applied to remove a total of five features and achieve an accuracy score of 40.4% on the validation set. While they certainly do not discriminate well on their own, these features capture information from the images, at a far cheaper computational cost and risk of overfitting than we would experience by using all pixel values from multiple filters.

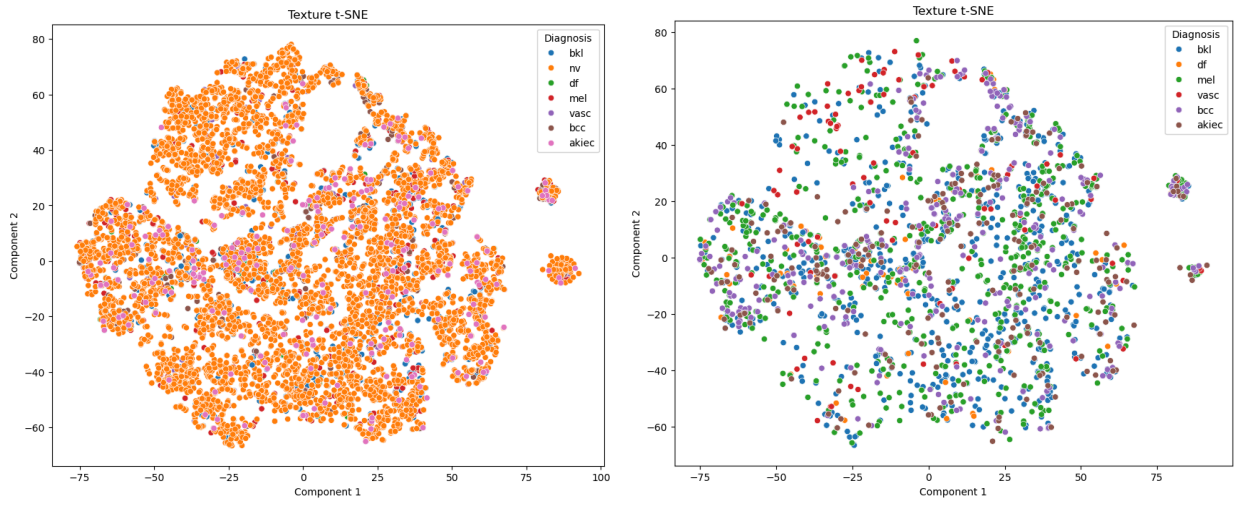


Figure 7. t-SNE of Texture

Using t-SNE to visualize if there is any separation of classes from the images, we noticed that all of the information appears to be clustered together. Even after removing NV, which is the dominant class, it appears that the feature doesn't provide great distinction between the different classes.

2.3 ResNet-50 with Principal Component Analysis

In an effort to retrieve features beyond our own feature extractions, we leveraged a pre-trained neural network, ResNet-50 (Residual Network-50), to output a feature vector. This network is a deep convolutional neural network (CNN) with 50 layers trained on millions of images and classes from the ImageNet database. To obtain a feature vector to be placed within our models, we input the original images that were in RGB, resized them to 244 by 244, and completed Principal Component Analysis (PCA) to reduce the dimensionality of the output.

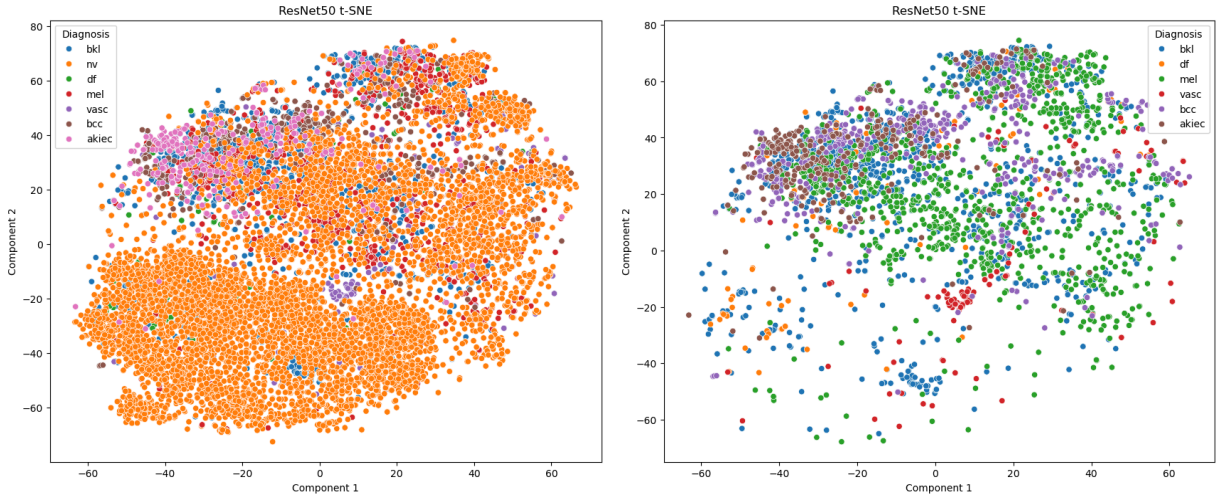


Figure 8. t-SNE of ResNet50

Using t-SNE to visualize if there is any separation of classes from the images, we noticed that all of the information appears to be clustered together. But after removing NV, there appears to be some clusters that are separated from other classes, such as AKEIC lesions. One potential explanation for the lack of separation is that ResNet-50 is trained to identify different objects rather than the same object with different characteristics. Therefore, these features may be less useful for our final model than we anticipated.

2.4 Pretrained ResNet on ImageNet Output as Features

In an effort to improve classification accuracy, we combined the output of ResNet. We passed the network's output through a max pooling layer, applied dropout with a rate of 0.5, passed the result through a 64-unit hidden layer with relu activation, then concatenated the result with our manually constructed features, and fed the result through another hidden layer, this time with 32 hidden units. Finally, the output was fed to a softmax layer for classification.

We observed no improvement after including these features. This is most likely due to the unique visual content of the images in our dataset, while public classification datasets focus on object recognition, our closeup medical images of skin lesions share little common visual features with those datasets. We believe complex features constructed using neural networks on this dataset or others like it only may yield better results, but leave that for further research since such efforts are outside the scope of this project.

3 Classification

We focused on two classification models we trained and analyzed, which are Random Forest and Logistic Regression. As discussed in Section 2, we used the following features: (1) metadata provided in the dataset, such as age, sex, and lesion location, (2) symmetry, (3) eccentricity, (4) mean HSV values, (5) texture statistical features, and (6) embeddings from the ResNet50 neural network model. There were a total of 792 features used for model training.

3.1 Results

For both classification models, we removed duplicate lesion ID images and split the dataset into 3 groups: (1) training set, representing 70% (5229 images) of the dataset, (2) validation set, representing 15% (1120 images) of the dataset, and (3) test set, representing 15% (1121 images) of the dataset. The training set for the Random Forest model was also upsampled using synthetic minority oversampling technique (SMOTE) to account for the class imbalance, totalling 26,474 samples. Training the Logistic Regression model using SMOTE was attempted but we struggled with convergence.

After initial model training, we use grid search to tune hyperparameters and use accuracy to evaluate performance. Figure 9 shows the hyperparameter(s) we tune and the best value(s). We then re-train our models using these hyperparameters and evaluate the performance on validation and test set. Figure 10 shows the classification accuracy of our final models.

Model	**Parameter**	**Search Values**	**Best Value**	**Best Accuracy**	**Search Time**
Random Forest	n_estimators	100, 200, 300	300	76.7%	8mins 42.3s
	max_depth	3, 5, 10, 12	12		
	class_weight	None, ‘balanced’, ‘balanced_subsample’	‘balanced’		
Logistic Regression	max_iter	100, 500, 1000, 2000	2000	75.7%	2mins
	class_weight	None, ‘balanced’	‘balanced’		

Figure 9. Summary of hyperparameters

Model	**Accuracy**	
	Validation Set	Test Set
Random Forest	77.6%	76.7%
Logistic Regression	72.5%	75.7%

Figure 10. Summary of results of final models

3.2 Model Performance

Both models performed well with the Random Forest model performing better on both the validation set and test set. With hyperparameter tuning, we saw uplifts in accuracy for both classifiers. The inclusion of embeddings from the ResNet50 neural network model did not improve model performance. The features extracted from ResNet50 might not be entirely relevant to the lesion classification task. This could be because ResNet50 is powerful for general image classification, but the features it learns might not align well with the nuances required for distinguishing between different types of lesions. The confusion matrices in Figures 12 and 13 provides a detailed overview of the performance by lesion type and highlights misclassifications.

4 Generalizability

Our logistic regression model generalizes well to the test dataset, in fact it happens to perform better on it than on the validation set. The random forest model achieved a slightly better score on both validation and test datasets. In both cases, the difference between in-sample and out-sample performance is not significant enough to indicate overfitting the validation set. This boosts our confidence in our hyperparameter selection and in our model's ability to generalize beyond this dataset to other images under the constraint that the images are captured in a similar manner.

5 Efficiency vs. Accuracy

The inclusion of ResNet features yielded results no better than our manually selected features. Further, of the two models we selected, efficiency is not a major concern. Although the random forest model shows an improvement of only 1% in test accuracy over the logistic regression model, the time it took to train was not significantly different between the two, as shown in the table below. The random forest took longer to optimize the hyperparameters compared to the logistic regression model, but at under nine minutes for a diagnosis tool, this is not a major limitation.

Model	Model Training Time	Prediction Time		
		Training Data	Validation Data	Testing Data
Logistic Regression	55s	0.5s	0.5s	0.5s
Random Forest	21.6s	0.1s	0.1s	0.1s

In terms of the features computation for training, the texture features' circles took about 10 hours to generate on 64 CPU cores. This is the most computationally intensive step in our feature pipeline, but we believe it can be improved with a better optimization algorithm. An attempt was made to improve efficiency by solving a linear regression on the circle's parameters fitted to canny-identified edges, but many of the resulting circles (as per example below) lacked precision. We opt for the grid-search approach and leave the algorithmic improvement to future works.

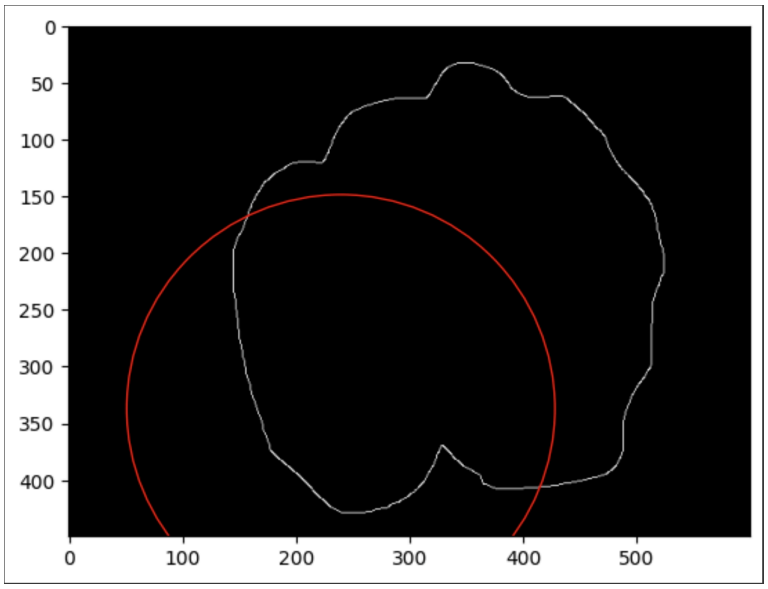


Figure 11.

6 Detailed Model Performance

Random Forest: Random Forest is an ensemble learning method that builds multiple decision trees and merges them to improve the overall prediction accuracy and control over-fitting. It handles high-dimensional data well and can capture non-linear relationships, making it well-suited for the lesion image classification task where the visual patterns can be complex and varied. Our Random Forest model performed well in aggregate. This was driven by the model's strong performance in classifying the "nv" class (F1 score of 0.88), which represented the most number of data points. However, the model had varying performance across the other classes.

Class	Precision	Recall	F1
akiec	0.30	0.40	0.34
bcc	0.42	0.57	0.49
bkl	0.44	0.41	0.43
df	0.25	0.36	0.30
mel	0.36	0.33	0.34
nv	0.90	0.89	0.89
vasc	0.63	0.5	0.56

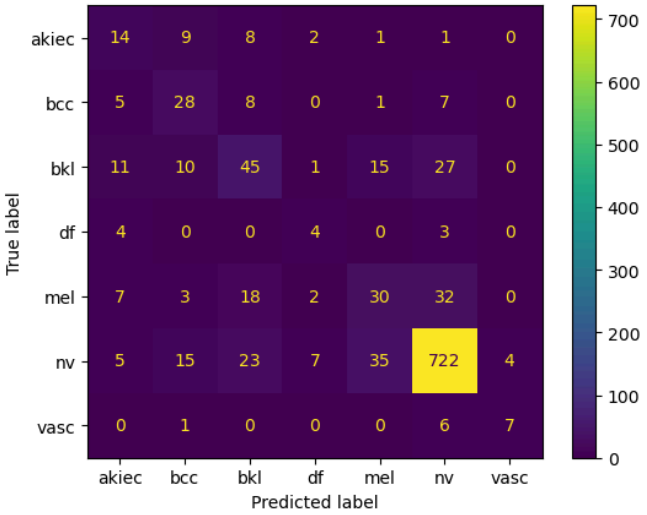
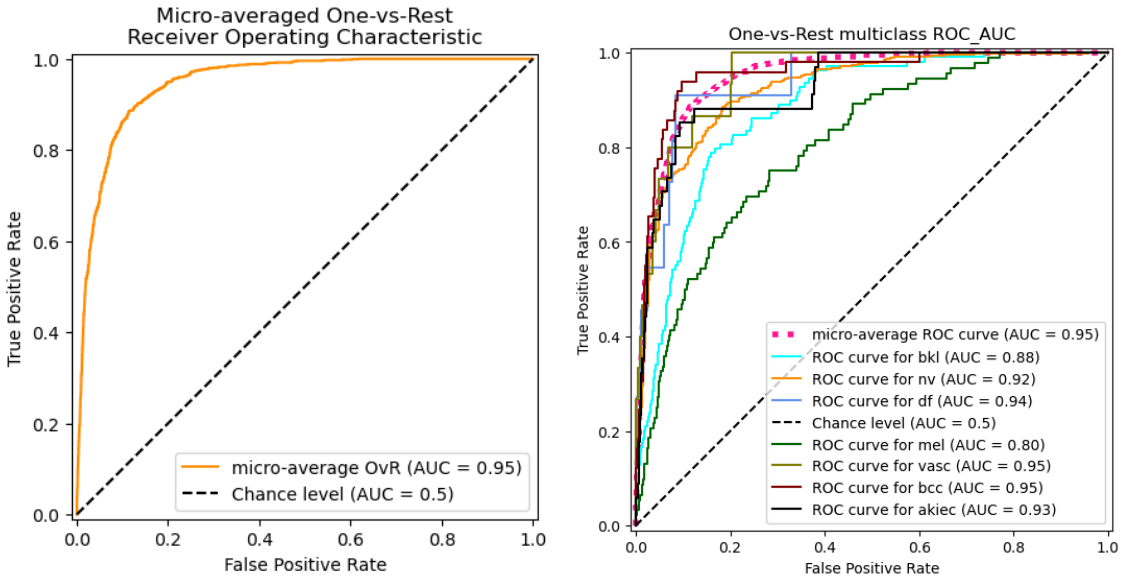


Figure 12. Confusion matrix for Random Forest model



Logistic Regression: Logistic Regression is a simple and easy to interpret classifier. It predicts the probability of each class based on one or more input features. It is suitable for problems where relationships between features and the target class are linear. Despite the challenges with obtaining complex relationships using Logistic Regression, our model performs well in aggregate. This was driven by the model's strong performance in classifying the "nv" class (F1 score of 0.88), which represented the most number of data points. However, similar to the Random Forest, the Logistic Regression model had varying performance across the other classes.

Class	Precision	Recall	F1
akiec	0.27	0.41	0.33
bcc	0.26	0.41	0.32
bkl	0.48	0.50	0.49
df	0.19	0.45	0.26
mel	0.36	0.59	0.44
nv	0.97	0.81	0.88
vasc	0.39	0.60	0.47

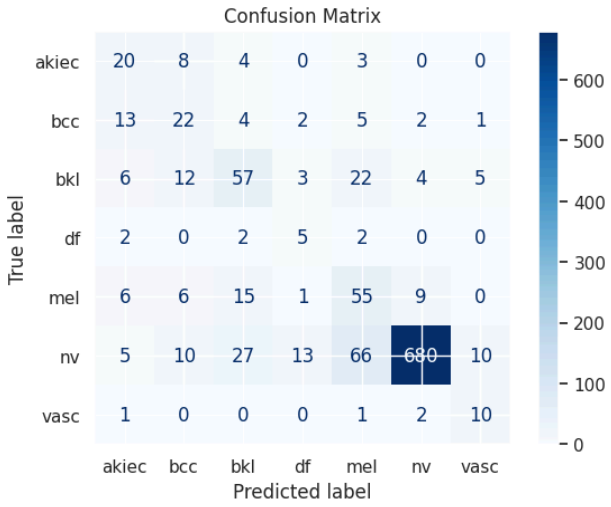
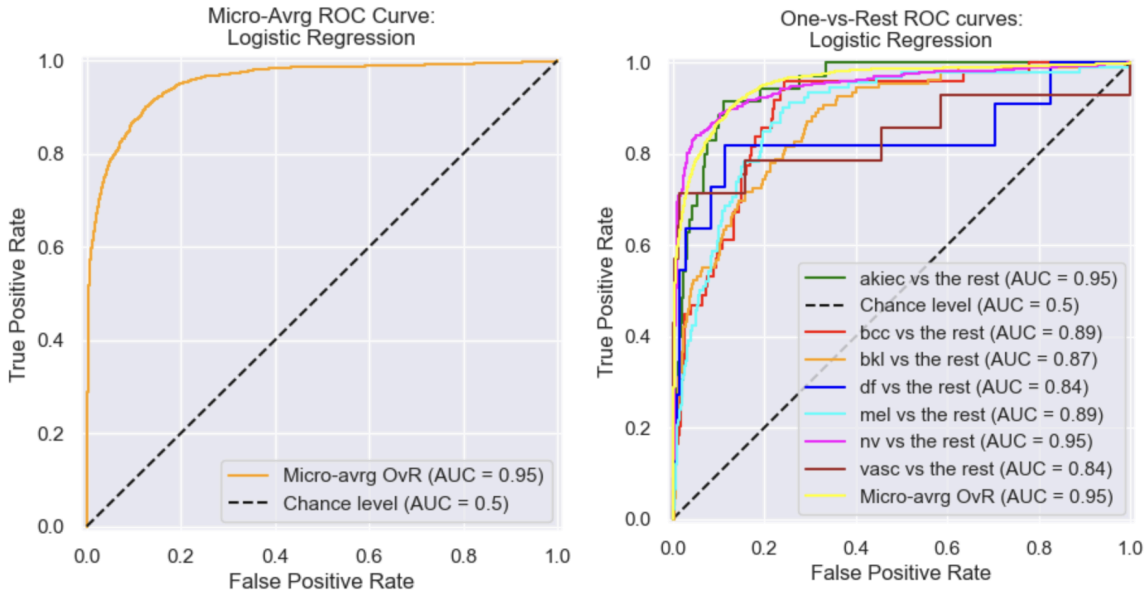


Figure 13. Confusion matrix for Logistic Regression model



7 Discussion

A major challenge was the class imbalance in the dataset, with an underrepresentation of cancerous classes compared to others. To address this, we applied class weighting to our models. This adjustment ensured that cancerous lesions, which were underrepresented, were given higher importance, preventing the models from being biased towards the noncancerous classes. This approach was crucial for enhancing the models' ability to detect cancerous lesions accurately. However, even with the class weighting and upsampling, our model was worse at predicting labels that are not nv. If possible, one of the next steps would be to obtain additional images of lesions of the minority class (DF and VASC).

Another limitation that resides in the NV class is that it appeared to cover characteristics of all the different types of lesions and was the majority class. This trend in the dataset makes sense as NV is what is commonly known as a mole. Noticing some irregularity in a mole usually leads to further biological testing, such as biopsies. For future steps, one approach we can take to classify the images better could be classifying them into cancerous and non-cancerous lesions and then classifying them into further subgroups of lesion types.

During the image pre-processing stage, we used the Dullrazor algorithm to remove hair and ruler marking within the image that impacted our ability to detect edges. One limitation of this method is that it can only detect dark hair. However, we did not notice an issue of edge detection with lighter color hair, therefore, this limitation minimally affects the feature extraction. Another limitation is that the algorithm sometimes removed blood vessels and other important color features within the images. This limitation may have affected the color extraction by decreasing the range of colors within the lesion, which may have impacted our ability to detect more granular differences between lesion types. Therefore, one future direction is to try a sharper hair removal method that only removes hair and ruler marking while leaving the other lesion structures in-tact. One potential next algorithm is SharpRazor[9].

We focused our dataset on features observable through visual inspection, however, our source data lacks information on the lesion-to-skin ratio to calculate area, as it consists only of close-up images of the lesion, and is limited to a two-dimensional view, preventing assessment of the lesion's protrusion above the skin. Consequently, we were constrained in developing a model that incorporates lesion size and skin protrusion. Additionally, the dataset does not provide information on the duration of lesion development, which restricts our ability to understand the lesion's progression over time.

Another future step is with the rise of tele-health, patients are taking pictures of their skin lesions on their phones and submitting it to clinicians to evaluate. It would be ideal to further train our system to generalize to images that are not taken in the same resolution and format as our dataset. This will further extend the capabilities of what this system can offer.

[9] *(Kasmi et al., 2023)*

8 Conclusion

Identifying whether skin lesions are cancerous by using visual clues is extremely important for dermatologists. In our project, we utilized the HAM10000 dataset to explore various techniques for improving the classification of these lesions. We focused on specific features such as symmetry, eccentricity, color, and texture, and assessed the performance of Random Forest and Logistic Regression models.

Despite some limitations in our feature extraction methods, the class imbalance, and the artifacts within the images, our models yielded encouraging results. The Random forest model, after hyperparameter tuning and applying class weights, showed the highest validation accuracy of 77.6%. This highlighted the effectiveness of combining simpler models with well-chosen features.

Given the importance of consistently correct diagnoses for the use of appropriate treatment, the accuracy we achieved is not quite at the level it can replace manual inspection or other testing methods, but suggests that with further research and availability to data, the tool can be made useful and help improve the care that can be provided to patients which limited access to medical professionals.

In terms of next steps and potential improvement, some ideas could be enhancing the precision of hair removal algorithms and developing more advanced feature extraction techniques could lead to better model performance. Another future step would be to increase the sample size of the minority classes and further develop the system to intake pictures that are not taken in a clinical setting.

In conclusion, our study demonstrates the potential of machine learning and computer vision in classifying skin lesions. Our goal is to assist healthcare professionals in making more accurate and timely diagnosis, ultimately benefiting patient outcomes.

References

Ali, A.-R. (2020, June 16). *Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images*. NCBI. Retrieved August 3, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7297317/

Halpern, A. C. (2021). *Melanoma Warning Signs and Images*. The Skin Cancer Foundation. Retrieved August 1, 2024, from https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/

How do I identify a suspicious lesion? (2020). Advanced Dermatology & Skin Surgery. Retrieved August 3, 2024, from https://advancedskindoctor.com/patient-resources/frequently-asked-questions/skin-cancer-frequently-asked-questions/how-do-i-identify-a-suspicious-lesion/

Tests For Melanoma Skin Cancer | Melanoma Diagnosis. (2023, October 27). American Cancer Society. Retrieved August 1, 2024, from https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/how-diagnosed.html

Philipp Tschandl (2018). *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. Harvard Dataverse. doi:https://doi.org/10.7910/dvn/dbw86t.

Tim Lee, Vincent Ng, Richard Gallagher, Andrew Coldman, David McLean, Dullrazor®: *A software approach to hair removal from images*, *Computers in Biology and Medicine*, Volume 27, Issue 6, 1997, Pages 33-543, https://doi.org/10.1016/S0010-4825(97)00020-6.

Kasmi R, Hagerty J, Young R, et al. *SharpRazor: Automatic removal of hair and ruler marks from dermoscopy images*. Skin Res Technol. 2023; 29:e13203. https://doi.org/10.1111/srt.13203