Skin Lesion Classification using Machine Learning

Aditya Gosavi
Department of Computer Science
Manipal Institute of Technology
Manipal, India
adityajgosavi@gmail.com

Kavish Shah
Department of Computer Science
Manipal Institute of Technology
Manipal, India
shahkavish0809@gmail.com

Shaurya Gupta
Department of Computer Science
Manipal Institute of Technology
Manipal, India

Kabir Bhalodi
Department of Computer Science
Manipal Institute of Technology
Manipal, India

***Abstract*— Skin cancer, a common and potentially fatal disease, benefits greatly from early detection for improved patient outcomes. Machine learning (ML) has shown promise in enhancing skin lesion classification accuracy and efficiency. This paper presents an approach using color, texture, and shape features from dermoscopic images in the HAM10000 dataset. Through data augmentation and feature extraction, including color histograms, Haralick textures, and Hu moments, we create a robust feature set. Using ensemble models like Decision Trees and Random Forest, we achieve high classification accuracy, highlighting ML's potential to support dermatologists in diagnostics.**

Keywords*— *Keywords—Skin Lesion Classification, Machine Learning, Skin Cancer Detection, HAM10000 Dataset, Feature Extraction, Color Histogram, Haralick Texture, Hu Moments, Data Augmentation, Ensemble Learning, Random Forest

I. Introduction

Skin cancer is one of the most common and potentially fatal forms of cancer globally, with incidences rising rapidly each year. Early detection and accurate classification of skin lesions are critical for effective treatment and improved patient outcomes. Conventional diagnostic methods, while effective, require specialized expertise and can be time-consuming. Consequently, there has been an increasing interest in leveraging machine learning techniques to aid in skin lesion classification, as these methods can provide faster, scalable, and more consistent results.

Machine learning algorithms have demonstrated significant potential in analyzing and classifying medical images by learning patterns from vast datasets. For skin lesion classification, various image features—such as color, texture, and shape—are vital, as these attributes often reflect the underlying pathological structure of skin lesions. The HAM10000 dataset, a large, diverse collection of dermatoscopic images, has been instrumental in advancing research in this area by providing a standardized benchmark for developing and testing machine learning models.

In this paper, we focus on classifying skin lesions into multiple categories using machine learning techniques. We use data augmentation to increase the dataset size, ensuring a balanced sample distribution for effective model training. We extract global features—such as color histograms for color, Hu Moments for shape, and Haralick Texture for texture—to create a comprehensive feature vector for each image. Ensemble learning models like Random Forest and Decision Trees are applied to classify the lesions, with a comparative evaluation conducted to determine the most effective model.
This research contributes to the field by offering a robust machine learning framework for skin lesion classification, emphasizing feature extraction and model optimization.

II. Dataset

The HAM10000 dataset, comprising 10,015 dermoscopic images of pigmented skin lesions, is the foundation of the proposed study. This dataset, sourced from multiple origins, is representative of common skin lesions but also exhibits significant class imbalances, a common challenge in medical datasets. The images, with a resolution of 600 × 450 pixels and saved in JPEG format, are cropped and centered around the lesions, with initial adjustments made for visual contrast and color reproduction. Each image is accompanied by seven attributes: patient age, patient sex, lesion ID (a unique identifier for each lesion type), image ID (a unique identifier for each image), diagnostic type for technical validation, the lesion's anatomical location, and the diagnostic category for skin lesions, which aids in diagnosing conditions. Further information about the dataset can be found in [1].

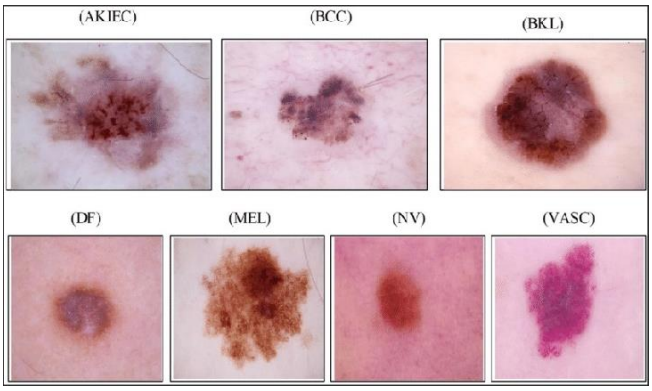

Figure 1. Sample images for the seven skin lesion classes from HAM10000 dataset.

The dataset also highlights significant skin cancer locations, such as the back, lower limbs, and trunk. The diagnostic skin lesion categories include seven distinct types, as shown in Figure 1, which presents sample images for each class.

The seven categories in the HAM10000 dataset are as follows:

1. **Actinic Keratoses [akiec]**: These are types of squamous cell carcinoma that are noninvasive and can typically be treated locally without surgery. There are 327 images available in the dataset for this category.
2. **Basal Cell Carcinoma [bcc]**: A type of epithelial skin cancer that rarely spreads but can be fatal if left untreated. The dataset includes 514 images for this category.
3. **Benign Keratosis-like Lesions [bkl]**: This category includes seborrheic keratoses, lichen-planus like keratoses, and solar lentigo, which are all considered benign lesions. There are 1,099 images for this category.
4. **Dermatofibroma [df]**: Skin lesions that are either benign growths or inflammatory responses to minor trauma. This category has 115 images in the dataset.
5. **Melanoma [mel]**: A cancerous tumor that develops from melanocytes. Melanoma can take various forms, but early detection allows it to be treated with a simple surgical procedure. The dataset includes 1,113 images for this category.
6. **Melanocytic Nevi [nv]**: Benign neoplasms of melanocytes, appearing in various shapes and sizes. This category has 6,705 images in the dataset.
7. **Vascular Lesions [vasc]**: Includes benign or malignant angiomas such as cherry angiomas, angiokeratomas, and pyogenic granulomas. There are 142 images available for this category.

Above information has been taken from [2].

III. METHODOLOGY

A. Data Preprocessing and Augmentation

- **Data Loading:** The images are loaded from the specified directory, and metadata (such as lesion types and image IDs) is extracted from the corresponding CSV file. The dataset includes images that were manually cropped around the lesion area, ensuring that each image represents the skin lesion clearly.

- **Data Augmentation:** Given the challenges of class imbalances in the dataset, horizontal flipping is employed to augment the data, effectively doubling the dataset size. Each image is flipped horizontally to generate additional samples for training the models, improving the generalization of the classifiers.

B. Feature Extraction

To capture key visual characteristics from the skin lesion images, several feature extraction techniques are applied:

- **Color Features**: The color histogram of each image is computed to capture the distribution of pixel intensities, which is important for identifying different lesion types based on color differences.

- **Texture Features**: Haralick texture features are extracted using the `mahotas` library, which quantifies texture patterns within an image and is particularly useful for distinguishing lesion types based on surface texture.
- **Shape Features**: Hu Moments are used to capture the shape of the lesions. These moments are invariant to image transformations such as scaling, rotation, and translation, making them robust features for shape classification.

C. Model Training and Classification

After experimenting with multiple machine learning models, three machine learning classifiers are employed to classify the skin lesions:

- **Decision Tree Classifier**: A decision tree model is trained on the extracted features, using the training dataset, and tested on the test dataset. It is capable of learning complex decision boundaries and provides interpretability for the classification process.

- **Random Forest Classifier**: To improve the decision-making process, an ensemble of decision trees is used. The Random Forest model aggregates predictions from multiple decision trees to increase classification accuracy and reduce overfitting.

- **K-Nearest Neighbors (KNN) Classifier:** The K-Nearest Neighbors (KNN) classifier is a simple, non-parametric algorithm used for classification tasks. For skin lesion classification, the KNN algorithm is trained by storing the feature vectors of the training dataset. During prediction, the algorithm calculates the distance between a test sample and all training samples, and the class of the test sample is determined by the majority vote of its k nearest neighbors.

D. Results and Evaluation Metrics

The evaluation metrics used are as follows:

- **Accuracy**: The accuracy of each model is computed, and a classification report is generated to evaluate precision, recall, and F1-score for each class.

- **Confusion Matrix**: A confusion matrix is plotted to visualize the model's performance in classifying different lesion types. It shows the true positive, false positive, true negative, and false negative counts for each class, helping to identify misclassifications.

- **Sample Visualization**: A set of random samples from the test set are displayed alongside their true and predicted labels, allowing visual inspection of the model's performance on individual cases.

The accuracies obtained are shown in Table 1.

Classifier Used	**Accuracy**
Decision Trees	0.9286
Random Forest	0.9501
KNN	0.6448

Table 1. Results of Machine Learning models

The confusion matrices are in Figure 2:

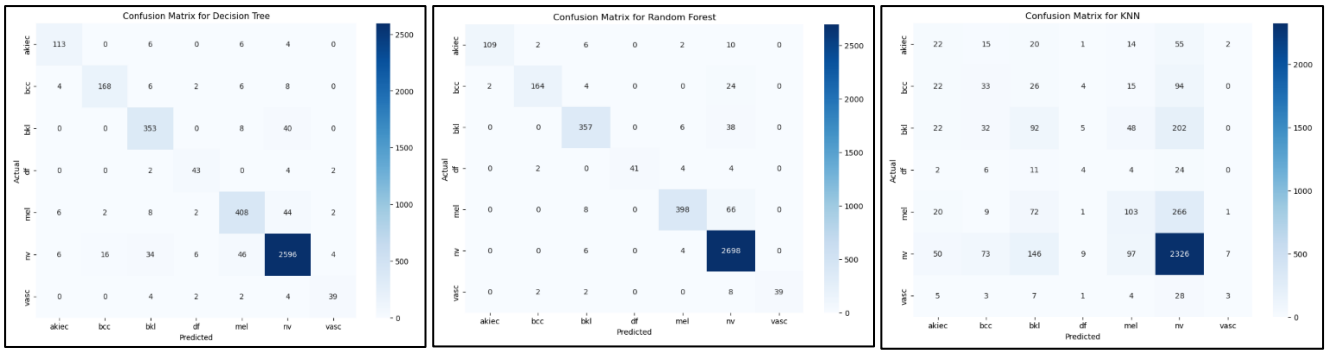

Figure 2. Confusion matrices for the models used.

The sample true and predicted labels with the images from the dataset are shown for the Random Forest model in Figure 3.

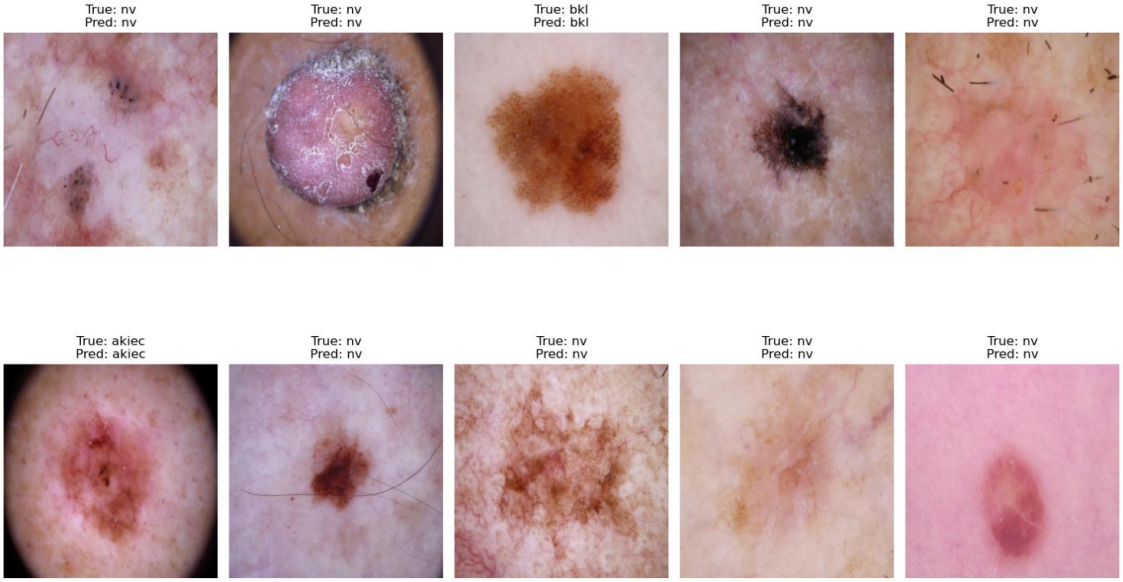

Figure 3. Sample Images with True and Predicated Labels.

The F1 Scores for each model and class are plotted and shown in Figure 4.

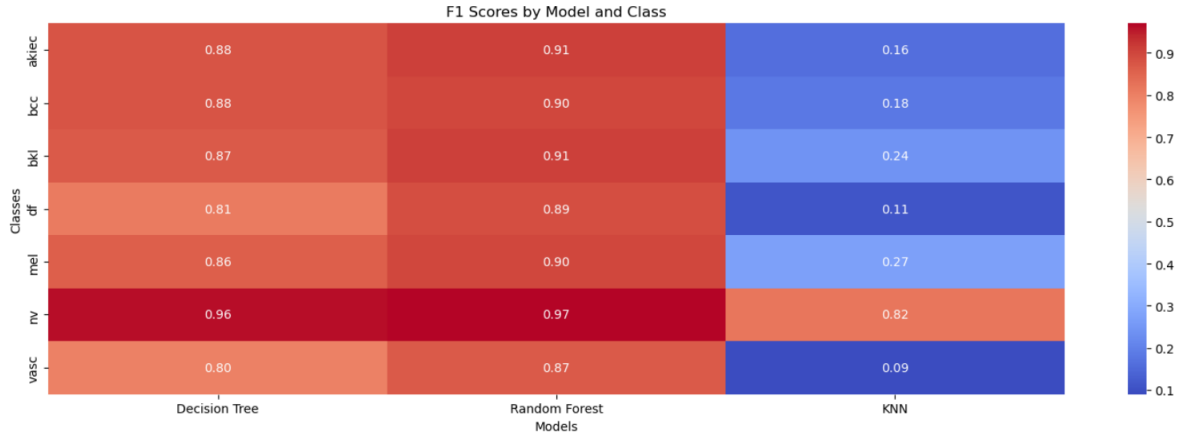

Figure 4. F1 Scores by Model and Class

RESULTS SUMMARY

Decision Tree Classifier:

- Accuracy: 92.86%
- Precision: High precision for most classes, particularly Melanocytic Nevi [nv] (96%), and relatively good precision for other categories like Basal Cell Carcinoma [bcc] (90%) and Benign Keratosis-like Lesions [bkl] (85%).
- Recall: High recall for Melanocytic Nevi [nv] (96%) and moderate recall for other common classes, with lower recall for rarer classes like Dermatofibroma [df] (84%) and Vascular Lesions [vasc] (76%).
- F1-Score: Balanced F1-scores across the categories, with a strong trade-off between precision and recall.
- Strengths: Simple, interpretable model with good accuracy and performance on well-represented classes.
- Weaknesses: Struggles slightly with rare classes due to class imbalance.

Random Forest Classifier:

- Accuracy: 95.01%
- Precision: Very high precision across most classes, with perfect precision for Vascular Lesions [vasc] (100%) and high precision for Melanocytic Nevi [nv] (95%).
- Recall: Excellent recall for Melanocytic Nevi [nv] (100%) but slightly lower for rare classes like Dermatofibroma [df] (80%) and Vascular Lesions [vasc] (76%).
- F1-Score: High F1-scores overall, showing a good balance between precision and recall.
- Strengths: Robust model with excellent handling of class imbalances, especially for rare categories, providing the highest accuracy and stable performance across all classes.
- Weaknesses: More complex and less interpretable than Decision Trees.

K-Nearest Neighbors (KNN) Classifier:

- Accuracy: 64.48%
- Precision: KNN struggled with precision, especially for rare categories such as Dermatofibroma [df] (11%) and Vascular Lesions [vasc] (9%). Melanocytic Nevi [nv] achieved a better precision (78%).
- Recall: Low recall for most classes, particularly for rare categories like Dermatofibroma and Vascular Lesions.
- F1-Score: Very low F1-scores, especially for rare classes, due to poor precision and recall.
- Strengths: Simple model, but struggles with class imbalances and high-dimensional data.
- Weaknesses: Poor overall performance with a significant drop in accuracy and recall, making it less suitable for this classification task, particularly with imbalanced classes.

Conclusion:

Random Forest performed the best with 95.01% accuracy, consistently handling both common and rare classes well, providing the highest precision, recall, and F1-scores.
Decision Tree also performed well with 92.86% accuracy, but it was less robust for rare categories compared to Random Forest.
K-Nearest Neighbors (KNN) performed poorly with 64.48% accuracy, struggling with both precision and recall for most classes, especially the rare ones, making it the least suitable model for this dataset.

Overall, Random Forest is the best performing model for skin lesion classification, followed by Decision Tree, while KNN should be reconsidered for this type of imbalanced dataset.

REFERENCES

[1] Tschandl, P. et al. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5:180161 doi: 10.1038/ sdata.2018.161 (2018).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Shetty B, Fernandes R, Rodrigues AP, Chengoden R, Bhattacharya S, Lakshmanna K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Sci Rep. 2022 Oct 28;12(1):18134. doi: 10.1038/s41598-022-22644-9. Erratum in: Sci Rep. 2022 Dec 19;12(1):21919. doi: 10.1038/s41598-022-26516-0. PMID: 36307467; PMCID: PMC9616944.