Skin Cancer Detection Using SVM with HOG Features and SMOTE Oversampling

Amrita School of Artificial Intelligence,
Amrita Vishwa Vidyapeetham, India
Himabala, Nikil, Sandeep, Chaithanya
{cb.sc.u4aie24028, cb.sc.u4aie24053, cb.sc.u4aie24047,
cb.sc.u4aie24040}cb.amrita.edu

***Abstract*—Skin cancer is one of the most common types of cancer, and its early diagnosis can significantly increase the chances of successful treatment. This paper presents an intelligent system for the detection of skin cancer using a Support Vector Machine (SVM) classifier, Histogram of Oriented Gradients (HOG) for feature extraction, and SMOTE for synthetic oversampling to handle class imbalance. We utilize the HAM10000 dataset, which includes diverse types of skin lesions. We describe the data preprocessing, feature extraction, model training, and evaluation phases in detail and also provide real-time detection capabilities using webcam input.**

***Index Terms*—Skin cancer, SVM, HOG, SMOTE, HAM10000, Image classification, Real-time detection, Machine learning, Computer vision, Dermatology**

I. Introduction

Skin cancer poses a significant healthcare burden due to its high prevalence and potential severity. The accurate and timely diagnosis of skin cancer can lead to effective treatments and improved patient outcomes. Dermatologists rely on dermatoscopic images and clinical experience to differentiate between benign and malignant skin lesions. However, this process is time-consuming and may be prone to human error.

To aid dermatologists and enhance diagnostic accuracy, we propose a machine learning-based system for skin cancer classification. Our approach combines effective image feature extraction using Histogram of Oriented Gradients (HOG), class balancing using Synthetic Minority Oversampling Technique (SMOTE), and classification using Support Vector Machines (SVMs). We also integrate a real-time webcam prediction module for practical use.

Our contribution includes:

- Use of traditional image processing with HOG instead of deep learning to reduce computational load.
- Custom implementation of SMOTE to synthetically balance minority classes.
- A robust evaluation protocol with accuracy, precision, recall, and F1-score.
- Deployment feasibility demonstration using webcam interface.

II. Related Work

Many previous studies have employed deep learning models such as Convolutional Neural Networks (CNNs) for skin lesion detection due to their superior performance. However, CNNs require significant computing resources and large datasets. Traditional machine learning models like SVMs are more efficient and interpretable. Few studies have combined HOG, SMOTE, and SVM for skin lesion classification, making our approach novel in its simplicity and practical deployability.

III. Dataset Description: HAM10000

The HAM10000 dataset, which stands for "Human Against Machine with 10000 training images," consists of 10,015 dermatoscopic images. These images are categorized into seven classes of skin lesions:

- **akiec** - Actinic keratoses and intraepithelial carcinoma (malignant)
- **bcc** - Basal cell carcinoma (malignant)
- **bkl** - Benign keratosis-like lesions (benign)
- **df** - Dermatofibroma (benign)
- **nv** - Melanocytic nevi (benign)
- **vasc** - Vascular lesions (benign)
- **mel** - Melanoma (malignant)

TABLE I: HAM10000 Class Distribution

Class	**Sample Count**
Melanocytic nevi (nv)	6705
Melanoma (mel)	1113
Benign keratosis-like lesions (bkl)	1099
Basal cell carcinoma (bcc)	514
Actinic keratoses (akiec)	327
Vascular lesions (vasc)	142
Dermatofibroma (df)	115

This distribution demonstrates a significant class imbalance that requires addressing for reliable classification results. For example, class ‘df‘ has 115 images while class ‘nv‘ has 6705, skewing training results if not corrected.

IV. Data Preprocessing

Before training the model, data preprocessing is applied. It includes:

- **Image resizing**: All images are resized to 64x64 for uniformity.
- **Grayscale conversion**: Simplifies image data for feature extraction.
- **Label encoding**: Converts textual class labels to numerical values.

- **Train-test split**: 80% data is used for training, 20% for testing.
- **Normalization**: Feature vectors are scaled between 0 and 1.

We use libraries such as NumPy, Pandas, scikit-image, and scikit-learn for preprocessing steps.

V. Feature Extraction using HOG

Histogram of Oriented Gradients (HOG) is a feature descriptor used to capture the structure and gradient orientation in images. It is widely used in image processing for object detection and classification due to its robustness and efficiency.

A. *Working of HOG*

The HOG feature extraction process includes:

1) **Grayscale Conversion:** Input images are converted to grayscale.
2) **Gradient Calculation:** Horizontal and vertical gradients are computed using Sobel filters.
3) **Orientation Binning:** Gradient directions are grouped into orientation histograms within each cell.
4) **Block Normalization:** Cells are grouped into overlapping blocks and normalized.
5) **Feature Vector:** All histograms are flattened into a single vector.

B. *Why HOG?*

- Captures edge and texture information effectively.
- Robust to variations in illumination.
- Computationally efficient for real-time applications.

We use ‘skimage.feature.hog()‘ from the scikit-image library with parameters:

- orientations=9
- pixels$_{p}er_{c}ell = (8, 8)cells_{p}er_{b}lock = (2, 2)$

VI. Synthetic Minority Oversampling Technique (SMOTE)

To address the imbalance in HAM10000, we employ SMOTE. It creates synthetic samples by interpolating between nearby minority class examples.

A. *Working Principle*

- For each minority class sample, select k-nearest neighbors.
- For a random neighbor, interpolate a new point:

$$x_{new} = x_i + \delta \cdot (x_{zi} - x_i)$$

where $\delta \in [0, 1]$.

B. *Benefits of SMOTE*

- Avoids overfitting.
- Balances dataset effectively.
- Enhances classifier performance on rare classes.
- Works better than random oversampling.

VII. Support Vector Machine Classifier

SVM finds the optimal hyperplane to separate classes in high-dimensional space.

A. *Model Training*

- SMOTE applied to training set only.
- One-vs-all approach for multi-class classification.
- Linear kernel chosen for simplicity and speed.
- Parameters tuned via grid search.
- Stratified K-Fold used for validation.

The classifier is implemented using ‘sklearn.svm.SVC‘. Training time is significantly reduced due to dimensionality reduction by HOG.

VIII. Evaluation and Results

In our skin cancer detection system, evaluating the performance of the model goes beyond just checking whether predictions are right or wrong. Since the dataset is **imbalanced** (some skin diseases have far more images than others), we use multiple metrics to understand how well the model is performing across all types of skin lesions. We use standard metrics:

- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision** = $\frac{TP}{TP+FP}$
- **Recall** = $\frac{TP}{TP+FN}$
- **F1 Score** = $2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

A. *Performance Metrics*

- Accuracy: 60.57%
- Macro Precision: 59.34%
- Macro Recall: 60.76%
- Macro F1 Score: 58.56%

We also generate a styled confusion matrix with ‘seaborn.heatmap()‘ for better visualization.

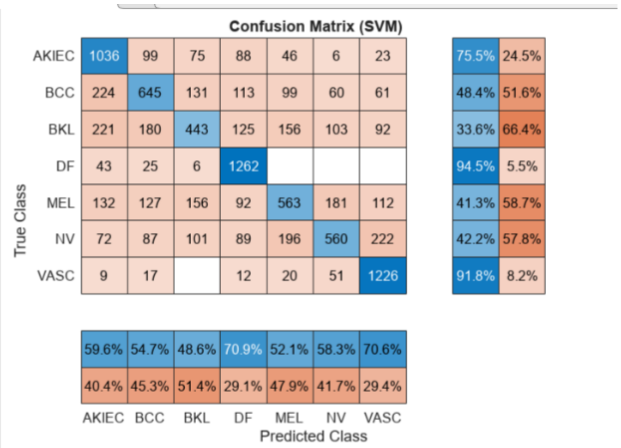

Fig. 1: Confusion Matrix of the SVM Classifier

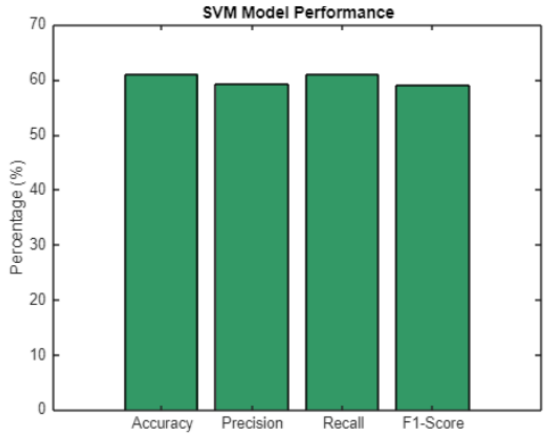

Fig. 2: Bar Graph Showing Evaluation Metrics

IX. Webcam-Based Real-Time Prediction

We implemented real-time image classification using a webcam. The workflow includes:

1) Capture frame from webcam.
2) Convert to grayscale.
3) Resize to 64x64.
4) Extract HOG features.
5) Predict using SVM classifier.

This enables interactive skin lesion analysis with instant results. We use OpenCV for video capture and scikit-learn for prediction.

X. Conclusion and Future Work

We propose a lightweight system using HOG, SMOTE, and SVM for skin lesion classification. Despite the simplicity, it achieves promising results. Future enhancements may include:

- Using deep learning models (e.g., CNNs) for better performance.
- Deploying on smartphones and edge devices.
- Incorporating explainable AI techniques.
- Expanding dataset for greater diversity.
- Hybrid models combining SVM with neural networks.

XI. References

1) Tschandl, P., Rosendahl, C., & Kittler, igH. (2018). The HAM10000 dataset. Scientific Data, 5, 180161.
2) Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. JAIR, 16, 321-357.
3) Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. CVPR.
4) Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
5) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. JMLR, 12, 2825–2830.
6) Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... & Yu, T. (2014). scikit-image: image processing in Python. PeerJ, 2, e453.
7) Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., ... & von Kalle, C. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer.