Diabetic Retinopathy Classification using Machine Learning based Techniques

Tabia Morshed
dept. of CSE
AUST
Dhaka, Bangladesh
tabia.cse.20200204027@aust.edu

Tanvir Md Raiyan
dept. of CSE
AUST
Dhaka, Bangladesh
tanvir.cse.20200204034@aust.edu

Musaddique Ali Erfan
dept. of CSE
AUST
Dhaka, Bangladesh
musaddique.cse.20200204049@aust.edu

***Abstract*—In this study, we present an image classification approach that evaluates multiple feature extraction techniques including Sobel edge detection, grayscale, color, and SIFT features. Using these extracted features, we train and test various classifiers such as K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Decision Tree to determine the most effective model for our dataset. Our results demonstrate that the Random Forest classifier consistently outperformed other methods, achieving the highest accuracy across most feature sets, particularly with grayscale and color features.**

***Index Terms*—Machine Learning, KNN, SVM, Decision Tree, Random Forest, Diabetic Retinopathy**

I. Introduction

Image classification is a fundamental problem in computer vision, involving the identification of objects or patterns within images. Various methods exist for extracting meaningful information from images, ranging from simple edge detection to more advanced techniques like SIFT (Scale-Invariant Feature Transform). Additionally, the effectiveness of classifiers plays a significant role in the performance of these image classification tasks. In this work, we explore the efficiency of different feature extraction methods—Sobel edge detection, grayscale, color, and SIFT—and compare the performance of KNN, Random Forest, SVM, and Decision Tree classifiers on a dataset of filtered images. By combining these techniques, we aim to find an optimal method for image classification in this domain.

II. Literature Review

[1] Yang et al. (2007) explore the effectiveness of bag-of-visual-words representations for scene classification by applying techniques from text categorization, such as term weighting, stop word removal, and feature selection, to generate image representations with different dimension, selection, and weighting of visual words. Their experiments demonstrate that binary visual-word features are as effective as tf or tf-idf weighted features and that frequent visual words are usually very informative and must not be removed.

[2] Jadhav et al. (2019) present a novel system for detecting and classifying soybean leaf diseases using color images, combining a multiclass support vector machine (SVM) and K-Nearest Neighbors (KNN) classifiers. The authors employ image processing and feature extraction techniques to classify three different types of diseases, achieving an accuracy of 87.3% for SVM and 83.6% for KNN. The system also incorporates a method for measuring disease severity based on lesion area.

[3] Nurwauziyah et al. (2018) compare the performance of decision tree, support vector machine (SVM), and k-Nearest Neighbor (KNN) classification methods using Pleiades and Landsat satellite imagery for land-use/cover mapping. The results indicate that SVM performed best in both high and low resolution imagery, achieving accuracies of 78.6% and 83.30%, respectively.

[4] Jasim and Al-Taei (2018) compare the performance of SVM and KNN algorithms for classifying plant leaf diseases based on visual features. Their results indicate that SVM achieved a higher overall accuracy (88.17%) than KNN (85.61%) and performed better on data with a higher degree of variation.

[5] Wasule and Sonar (2017) propose a system for classifying brain MRI images into malignant vs. benign and low-grade vs. high-grade glioma using GLCM features and SVM and KNN classifiers. Their results show that the proposed system achieves accuracies of 96% for SVM and 86% for KNN on a clinical database and 85% for SVM and 72.50% for KNN on a Brats database, demonstrating the effectiveness of these classification approaches for brain tumor identification.

III. Methodology

A. Dataset

The dataset consists of Gaussian-filtered images of varying stages of a medical condition, divided into five categories: No_DR, Mild, Moderate, Severe, and Proliferate_DR. The dataset is accompanied by a CSV file containing the image filenames and their corresponding labels.

B. Feature Extraction

We extracted four sets of features from the images:

- Sobel Edge Detection: This method highlights the edges in an image using the Sobel operator to capture the gradient in both horizontal and vertical directions. The resulting edge maps are resized to 28x28 pixels and flattened.

- Grayscale Features: The images are converted to grayscale and resized to 28x28 pixels, then flattened into 1D feature vectors.
- Color Features: Images are converted to the HSV color space, and the mean and standard deviation of the H, S, and V channels are computed to form a six-dimensional feature vector.
- SIFT Features: Keypoints and descriptors are extracted using the SIFT algorithm. A Bag of Visual Words (BoVW) model is created using k-means clustering to reduce the dimensionality and build a histogram of visual words for each image.

C. Models

Four classifiers were employed to evaluate the performance of the extracted features:

- K-Nearest Neighbors (KNN): A simple algorithm that classifies a sample based on the majority label of its nearest neighbors.
- Random Forest: An ensemble method that builds multiple decision trees and aggregates their predictions to improve accuracy.
- Support Vector Machine (SVM): A robust classifier that aims to find the optimal hyperplane that maximizes the margin between different classes.
- Decision Tree: A simple and interpretable model that splits data based on the most informative features at each node.

D. Evaluation

We split the dataset into 80% training and 20% testing sets. Each classifier was trained using the different feature sets, and the classification accuracy was calculated for each model.

IV. Result Analysis

The classification results for each feature set are as follows:

TABLE I
PERFORMANCE OF MACHINE LEARNING MODELS

Model	KNN	RF	SVM	DT
Sobel	0.702	0.736	0.729	0.635
SIFT	0.645	0.664	0.690	0.579
Color	0.693	0.733	0.702	0.650
Grayscale	0.698	0.740	0.724	0.653

The results suggest that Random Forest is the most robust classifier across different feature extraction methods, particularly excelling with grayscale and color features. Meanwhile, SVM performed well with SIFT features, while KNN showed consistent results across all feature sets.

References

[1] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in Proc. of the 2007 ACM Int'l Conf. on Image and Video Retrieval (MIR '07), pp. 197-206, Sept. 2007.

[2] S. B. Jadhav, V. R. Udupi, and S. B. Patil, "Soybean leaf disease detection and severity measurement using multiclass SVM and KNN classifier," Int. J. Elec. & Comp. Eng., vol. 9, no. 5, pp. 4077–4091, Oct. 2019.

[3] I. Nurwauziyah, U. D. Sulistyah, I. G. B. Putra, and M. I. Firdaus, "Satellite image classification using decision tree, SVM and k-nearest neighbor," ResearchGate, Jul. 2018.

[4] S. S. Jasim and A. A. M. Al-Taei, "A Comparison Between SVM and K-NN for Classification of Plant Diseases," DIYALA J. Pure Sci., vol. 14, no. 2, pp. 94–105, Apr. 2018.

[5] V. Wasule and P. Sonar, "Classification of brain MRI using SVM and KNN classifier," in Proc. 2017 IEEE 3rd Int. Conf. Sens., Signal Process. Secur. (ICSSS), pp. 218-223, Sept. 2017.