Original Research Article

Banana Disease Diagnosis Using Hybrid Machine Learning Models

ABSTRACT

Diseases and pests pose a serious threat to the agricultural sector, significantly reducing both production and economic viability. Banana plants are particularly vulnerable to a range of pests and diseases, which can substantially impact yield and quality if not effectively managed. To address this issue, the integration of machine learning paradigms can facilitate the early detection of diseases. Specifically, the combination of region-based active contours and K-Nearest Neighbours (KNN) classification offers a comprehensive approach to object detection and classification in images. This integrated method leverages the strengths of both techniques, enabling accurate object segmentation through active contours and effective classification based on extracted features using K-Nearest Neighbours. This approach promises to enhance disease detection in banana plants, thereby improving agricultural productivity and economic outcomes.

Keywords: Chan-Vese, Histogram Equalizer, k-Nearest Neighbour (KNN), Principal Component Analysis (PCA)

1 INTRODUCTION

Region-Based Active Contour (RBAC) models, also known as region-based segmentation models, have become a fundamental tool in image processing and computer vision. These models are designed to detect and delineate the boundaries of objects within an image by evolving a contour or curve based on region information [1]. Unlike edge-based methods that rely heavily on gradient information and are sensitive to noise and weak boundaries, RBAC models utilize the statistical information of regions to achieve more robust segmentation results.

The core idea of RBAC is to minimize an energy functional that integrates the differences between the statistical properties of the regions inside and outside the contour. Typically, these properties include intensity, texture, and color. By leveraging this regional information, RBAC models can effectively handle images with intensity inhomogeneity, complex structures, and occlusions [1].

One of the pioneering approaches in RBAC is the Chan-Vese model, which formulates the segmentation problem as a variation problem [2][3]. This model introduces a level set method to represent the evolving contour implicitly, allowing for topological changes such as splitting and merging. The Chan-Vese model minimizes an energy functional that is

composed of data fidelity and regularization terms, making it a powerful and flexible tool for various image segmentation tasks [2].

Over the years, RBAC models have been extended and refined to incorporate additional information and constraints, leading to improved performance in diverse applications such as medical imaging, object tracking, and scene understanding. These advancements have solidified RBAC as a versatile and essential technique in the toolbox of modern image analysis. Region-Based Active Contour models have emerged as a pivotal tool in the realm of plant disease analysis, significantly enhancing the accuracy and efficiency of disease detection and monitoring [2]. These models leverage regional information within images to delineate diseased areas, providing a robust solution to the challenges posed by the complex visual characteristics of plant diseases.

Plant disease detection often involves analysing leaf images where symptoms such as spots, discolorations, or texture changes need to be accurately identified and segmented. Traditional methods relying on edge detection or manual inspection are often insufficient due to varying intensity, shape, and spread of disease symptoms. RBAC models address these challenges by utilizing the statistical properties of the image regions to evolve a contour that precisely encloses the affected areas [2][3]. A notable application of RBAC in plant disease analysis is the segmentation of lesions caused by bacteria, fungal, or viral infections. By minimizing an energy functional that accounts for the intensity and texture differences between healthy and diseased regions, RBAC models can accurately isolate the symptomatic areas, even in the presence of noise or uneven illumination. This capability is particularly beneficial for early disease detection, where subtle symptoms need to be identified before they become visually obvious. RBAC models have not only improved the speed and accuracy of disease detection but have also contributed to more sustainable and effective plant health management practices.

2 LITERATURE REVIEW

This section discusses the various works as carried out by different researchers that are related to our study.

Rahaouma et. al [10] in their study introduces a computer-aided detection (CAD) system using computed tomography (CT) scans for lung nodule classification, incorporating image processing, segmentation, feature extraction via Discrete Wavelet Transform (DWT) and classification using Polynomial Neural Network (PNN), achieving a high accuracy of 96.66%.

The study made by Radhi and Kamil [11] compares snakes and level sets for breast tumour segmentation in mammograms, with Chan-Vese showing superior performance (90%, 95%, 98%, 97%, and 97%) in Jaccard, Dice, PF-Score, Precision, and Sensitivity metrics, highlighting its reliability for computer-assisted detection systems.

Deriche et. al [12] in their paper introduces a robust image segmentation method combining convex active contours with the Chan-Vese model, minimizing user input dependency and enhancing segmentation accuracy across diverse image types. Experimental results demonstrate superior performance in both processing time and segmentation accuracy compared to recent methods across standard image databases.

Ali et. al [13] in their paper introduces a method ΔE color difference and texture features for automatic detection and classification of citrus diseases, achieving 99.9% accuracy and

comparable sensitivity, validated by Area Under Curve (AUC) of 0.99. Feature reduction via Principal Component Analysis (PCA) and testing with advanced classifiers further substantiates the robustness of the proposed approach.

Zhang et. al [14] in their paper demonstrates the efficacy of PCA and cluster analysis for distinguishing late blight infected tomatoes from healthy ones based on spectral characteristics. It highlights the potential of spectral ratio analysis to identify sensitive wavelengths critical for accurate disease detection in remote sensing applications.

Füzy et. al [15] in their study uses PCA to identify key indicators of drought and salt stress in plants, highlighting parameters such as root electrical capacitance, membrane stability index, leaf relative water content, and SPAD units as sensitive stress indicators across diverse experimental setups.

Amato and Falchi [16] in their study introduces a novel kNN (k Nearest Neighbour) based image classification method focusing on local feature similarity, enhancing efficiency and effectiveness in landmark recognition tasks using various types of local features.

3 MATERIAL AND METHODS

3.1 Histogram Equalizer

Histogram equalization is a technique in image processing used to improve the contrast of an image [5]. This method can be particularly useful in plant disease classification, where high-contrast images can enhance the visibility of disease symptoms on plant leaves, stems, or fruits. Histogram equalization works by redistributing the intensity values of an image so that they span the entire range of possible values [4]. This process makes the image details more visible and improves the overall contrast. The histogram of the image is computed which shows the frequency of each intensity value. For representing the cumulative sum of the histogram values, the Cumulative Distribution Function (CDF) is calculated. The CDF is normalized so that its value ranges from 0 to 1 [4]. The normalized CDF is utilized to map the original intensity values to new values, resulting in a contrast-enhanced image. Histogram equalization is a valuable preprocessing step in the context of plant disease classification. By improving the contrast of plant images, it facilitates the extraction of more distinct features, thereby enhancing the accuracy of disease detection and classification. This technique, combined with robust classification algorithms, can significantly contribute to the effective management and diagnosis of plant diseases.

3.2 Chan – Vese algorithm

The Chan-Vese algorithm is a method used in image processing and computer vision for image segmentation. It is based on the level set method and is particularly effective for segmenting objects in images with weak or missing boundaries. It segments an image into regions, typically foreground and background, based on their intensities [2]. It is built on the Mumford-Shah functional for segmentation and utilizes the level set method to represent the contour. In the Chan-Vese algorithm, the evolving contour is represented as the zero-level set of a higher-dimensional function [2]. The algorithm aims to minimize an energy functional that depends on the contour and the image data. This energy functional typically has terms that represent the difference in intensities inside and outside the contour, as well as regularization terms to ensure smoothness. The process continues until the contour stabilizes, indicating that the optimal segmentation has been found.

3.3 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of data while retaining most of the variation present in the database [6]. In the context of plant disease detection and classification, PCA can be employed to simplify complex image data, making it easier to process and analyze while preserving critical information about disease symptoms. PCA transforms the original data into a new coordinate system, where the greatest variance by any projection of the data comes to lie on the first coordinate (the first principal component), the second greatest variance on the second coordinate, and so on. This transformation allows for a reduced representation of the data that captures the most important features [6]. Firstly, it should be ensured that each feature has a mean of zero and a standard deviation of one. The covariance matrix is computed to understand the relationships between different features. Eigenvalues and eigenvectors of the covariance matrix are calculated. The eigenvectors represent the principal components, and the eigenvalues indicate the amount of variance captured by each principal component. A feature vector is formed by selecting the top 'k' eigenvectors based on their corresponding eigenvalues. The original data is projected onto the new feature space using the feature vector, resulting in a reduced dataset. By reducing the complexity of the feature set, PCA helps in focusing on the most significant features, thereby improving the efficiency and effectiveness of classification algorithms [6].

3.4 Standard Scaler

Standard scaling, also known as Z-score normalization, is a preprocessing technique used to standardize the features of a dataset. In the context of plant disease detection and classification, applying a standard scaler helps to normalize the features extracted from plant images, ensuring that each feature contributes equally to the analysis [7]. Standard scaling transforms the data such that it has a mean of zero and a standard deviation of one. This ensures that all features are on a comparable scale, which is particularly important for algorithms sensitive to feature scales, such as k-Nearest Neighbours (KNN) or Support Vector Machines (SVM). The working of Standard Scaler is as follows. At first the mean (μ) and standard deviation (σ) for each feature in the dataset is computed [7]. Subtract the mean and divide by the standard deviation for each feature [7].

$$x` = x - \mu \tag{1}$$

Where x is the original feature value, and x` is the standardized feature value.

3.5 K-Nearest Neighbour

The k-Nearest Neighbours (KNN) algorithm is a simple, yet powerful, supervised machine learning algorithm that is used for both classification and regression tasks [8]. KNN is based on the concept of similarity or distance between data points [9]. KNN operates on the principle that similar instances exist near each other. The algorithm identifies the 'k' closest data points to a given test instance and assigns the most common label (for classification) or the average value (for regression) among these neighbours to the test instance [8]. Various distance metrics like Euclidean, Manhattan, or Minkowski distances are used to measure the closeness between data points. For classification, the majority class among the neighbours is assigned to the test instance. For regression, the mean value of the neighbours is assigned. In KNN, the entire dataset is used during testing [9].

The KNN algorithm is a widely used machine learning technique for classification tasks. Its simplicity and effectiveness make it particularly useful in fields such as plant disease

classification, where the goal is to categorize plant conditions based on visual symptoms. It can classify different diseases affecting plants by analysing features extracted from images of plant leaves, stems, or fruits. These features may include color, texture, shape, and other visual characteristics. KNN is easy to understand and implement, making it accessible for researchers and practitioners. KNN can perform well even with relatively small datasets, which is often the case in plant diseases classification [8]. Plant disease datasets are often imbalanced, with some diseases being underrepresented. This can affect the accuracy of KNN, which may require techniques like oversampling or synthetic data generation.

3.6 Methodology

The dataset comprises ten thousand images of banana plants affected by seven different diseases and pests. The classes include Black Sigatoka, Yellow Sigatoka, Panama disease, Pseudostem Weevil, Banana Aphids, Scarring Beetle, and Bacterial Soft Rot. All images are initially converted to grayscale to reduce computational complexity and focus on intensity variations which are essential for subsequent processing steps. Histogram equalization technique is applied to the grayscale images to improve the contrast. The equalizer redistributes the intensity values of the pixels to span the entire range, enhancing the visual contrast and highlighting features that are important for disease and pest identification.

Chan-Vese Region-Based Active Contour Segmentation method is used to segment the histogram equalized images. The Chan-Vese model is effective for segmenting objects in images where the boundaries are not well-defined. It operates by evolving a contour to partition the image into regions, minimizing the difference in intensities within each region. This helps in isolating the affected areas from the healthy parts of the banana plant in the images.

PCA is employed to reduce the dimensionality of the feature space. By transforming the original high-dimensional data into a lower-dimensional form, PCA helps in capturing the most significant features while discarding redundant information. Texture features, which capture the surface characteristics and patterns of the banana diseases and pests, are crucial for distinguishing between the different classes of diseases and pests based on their visual appearance.

The KNN algorithm is used to classify the images into the seven categories of banana diseases and pests. KNN is a simple, yet effective, classification method that assigns a class to an image based on the majority class among its k-nearest neighbours in the feature space. The distance metric, usually Euclidean, is used to determine the nearest neighbours, and the value of k is chosen based on cross-validation to optimize the classifier's performance.

3.6.1 Evaluation Metrics

The performance of the proposed model was assessed using a range of widely recognized and relevant evaluation metrics. Specifically, the model was evaluated based on standard metrics including precision, recall, and F1-Score, which are commonly used to measure classification effectiveness in various domains.

- Precision – It measures the accuracy of positive predictions by calculating the ratio of true positive predictions to the total number of positive predictions.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

High precision means the model predicted correctly.

- Recall – It measures the relevant positive instances by calculating the ratio between true positive predictions to the total number of actual positive instances.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

- F1-Score – F1-Score is the harmonic mean of precision and recall. It balances both precision and recall.

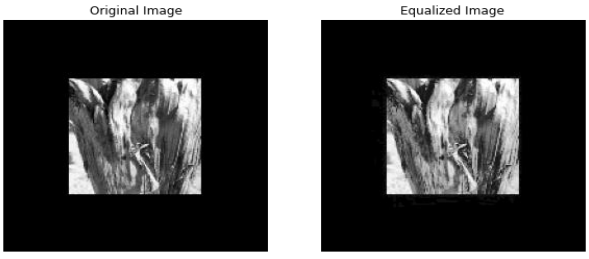
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

High F1-Score indicates a good balance between Precision and Recall.

4. RESULTS AND DISCUSSION

This section details the results of each step in the methodology, highlighting the effectiveness of the preprocessing, segmentation, dimensionality reduction, and classification techniques used in this study.
The first step in our workflow was to convert the RGB images into grayscale images. This conversion simplifies the data by focusing on intensity variations, which are crucial for detecting disease symptoms and pests on banana plants. Histogram equalization was applied to the grayscale images to enhance contrast, as illustrated in figure 1(C). This step distributed the intensity values, making important features more distinguishable.

(A)



(B) (C)

Fig. 1. Sample images (A) exhibiting the original image, (B) Grayscale image, and (C) the Histogram Equalized Image

Following histogram equalization, the Chan-Vese segmentation algorithm was applied to the images. The Chan-Vese method is effective for segmenting objects where boundaries are not well-defined, which is essential for isolating affected areas from healthy parts of the banana plants. The segmented images are presented in figure 2(E).

(D) (E)

Fig. 2. Sample images (D) exhibiting the Equalized image, (E) Chan-Vese segmented Image

Principal Component Analysis (PCA) was employed on the combined dataset to reduce its dimensionality to 2 for visualization purposes. A scatter plot of the principal components was generated, revealing the distribution of images in the reduced space which is shown in figure 3. Additionally, a heatmap of PCA was created after applying Z-score normalization and hierarchical clustering, as shown in figure 4.

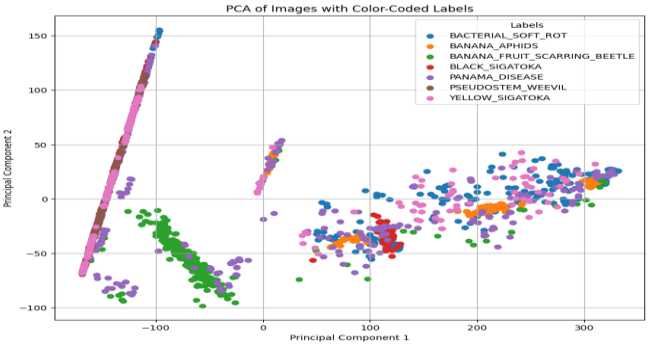


Fig. 3. Scatter plot of Principal Component Analysis of images with color-coded labels.

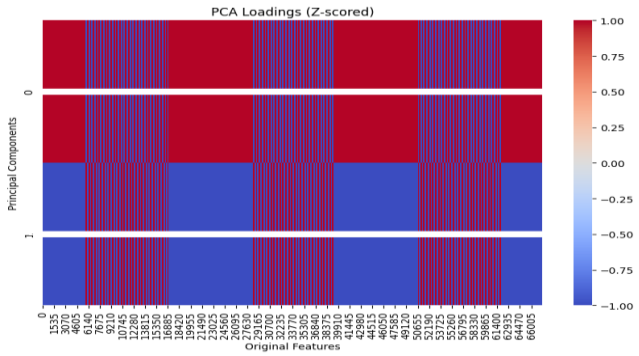


Fig. 4. Heat map of Principal Component Analysis after applying Z-score of normalization.

The data was standardized using StandardScaler before PCA. The heatmap visualization provided valuable insights into the structure and relationships within the dataset, showing how the original features contributed to the principal components and the variance explained by each component.

A KNN classifier was used to classify the features extracted by the PCA feature extractor. The KNN classifier was instantiated with 2 neighbours. Upon training the model, various parametric metrics were generated. The confusion matrix, displayed in figure 5, provides a clear view of the models' performance in terms of true positives, true negatives, false positives, and false negatives.

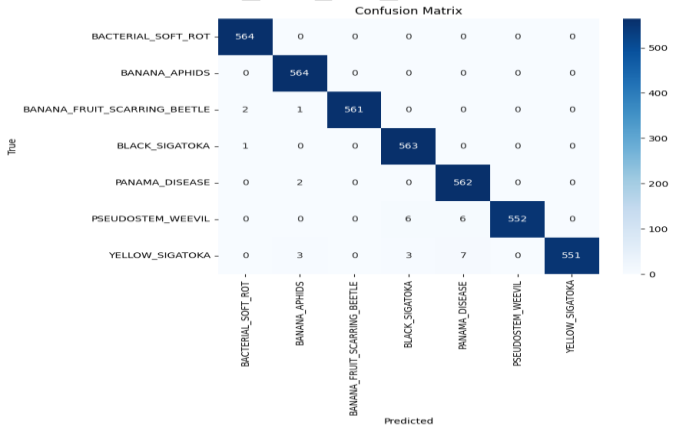


Fig. 5. Confusion matrix generated from the K-Nearest Neighbour classifier.

To evaluate the robustness of the model, k-Fold Cross-Validation was applied, dividing the data into k equally sized folds. The model was trained on k-1 folds and tested on the remaining fold. This process was repeated k times, with each fold used exactly once as the test set. The final performance metric was the average of the metrics obtained from each fold.

Cross-Validation Scores: [0.97979798 0.95959596 0.96464646 0.93939394 0.93434343 0.94444444 0.95959596 0.93939394 0.88832487 0.64974619 0.62944162 0.97461929 0.96954315 0.95939086 0.93908629 0.94416244 0.93908629 0.95939086 0.94923858 0.9035533]

Mean Accuracy: 0.9163397938778649

The cross-validation scores, representing the accuracy of the model on each fold, indicated a mean accuracy of 91.63%. Our array suggested a k-fold cross-validation with 20 folds, as there were 20 performance metrics listed. The data was divided into 20 parts, each part used once as a validation set while the remaining 19 parts were used for training in each iteration. The average accuracy of 91.63% provides an overall performance estimate for the model, demonstrating its effectiveness in classifying banana diseases and pests.

The ROC (Receiver Operating Characteristic) curve illustrated in figure 6, was calculated to evaluate the performance of the KNN classifier in terms of its ability to discriminate between different classes. Our model demonstrates a strong ROC curve during classification. The curve shows a high True Positive Rate (Sensitivity), indicating that our model is performing well and accurately identifying the positive cases.

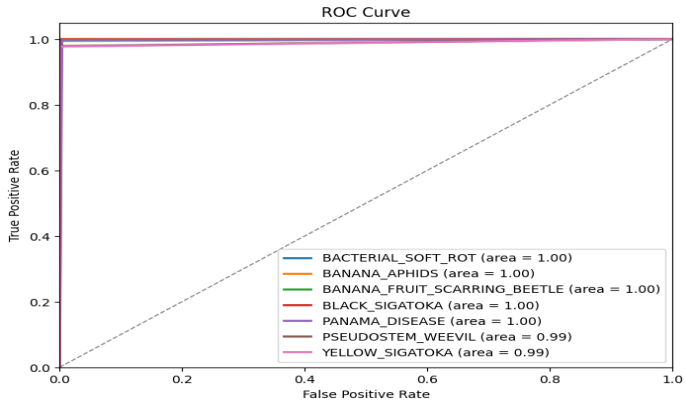


Fig. 6. Generation of Receiver Operating Characteristics curve after classification.

From the model we can indicate that a high precision is accurate for most of the predicted class. False positives are relatively rare, and the model does not incorrectly label regions. The high recall value indicates that the model identifies all relevant instances of the target class. In the context of active contour models, the model successfully identifies most of the important contours without missing them. There is also a good balance between the precision and recall. The model is both accurate and comprehensive. The values of the evaluated matrices are given in table 1.

Table 1 Classification report of modified region based active contour model.

CLASS	PRECISION	RECALL	F1-SCORE
APHID	0.99	1.00	0.99
B_SIGATOKA	0.98	1.00	0.99
PANAMA	0.98	1.00	0.99
SOFT_ROT	0.99	1.00	1.00
S_BEETLE	1.00	0.99	1.00
WEEVIL	1.00	0.98	0.99
Y_SIGATOKA	1.00	0.98	0.99

5. CONCLUSION

The proposed methodology offers a systematic approach to image classification, from preprocessing and feature extraction to model training and evaluation. By incorporating advanced techniques such as region-based active contour transformation and dimensionality reduction, it aims to enhance the accuracy and interpretability of classification results. Additionally, the use of performance evaluation metrics ensures a robust assessment of the classifier's performance, enabling informed decision-making in real-word applications.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declares that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text to image generators have been used during writing or editing of this manuscript.

REFERENCES

[1] Appia V, Yezzi A. Active geodesics: Region-based active contour segmentation with a global edge-based constraint, 2011 International Conference on Computer Vision. 2011;1975-1980.

[2] Saini K, Dewal ML, Rohit M. A fast region-based active contour model for boundary detection of echocardiographic images. Journal of digital imaging. 2012;25: 271-278.

[3] Wong OQ, Rajendran P. Image segmentation sing modified region-based active contour model. J. Eng. Appl. Sci. 2019;14: 5710-5718.

[4] Garg P, Jain T. A comparative study on histogram equalization and cumulative histogram equalization. International Journal of New Technology and Research. 2017;3(9): 263242.

[5] Cheng HD, Shi XJ. A simple and effective histogram equalization approach to image enhancement. Digital signal processing. 2004;14(2): 158-170.

[6] Li L, Liu S, Peng Y, Sun Z. Overview of principal component analysis algorithm. Optik. 2016;127(9): 3935-3944.

[7] Thara DK, PremaSudha BG, Xiong F. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognition Letters. 2019;128: 544-550.

[8] Pandey A, Jain A. Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security. 2017;10(11): 36.

[9] Fan Z, Xie JK, Wang ZY, Liu PC, Qu SJ, Huo L. Image classification method based on improved KNN algorithm. 2021 Journal of physics: Conference series. 2021;1930(1): 012009.

[10] Rahouma KH, Mabrouk SM, Aouf M. Lung cancer diagnosis based on chan-vese active contour and polynomial neural network. Procedia Computer Science. 2021;194: 22-31.

[11] Radhi EA, Kamil MY. Breast Tumor Detection Via Active Contour Technique. International Journal of Intelligent Engineering & Systems. 2021;14(4).

[12] Deriche M, Amin A, Qureshi M. Color image segmentation by combining the convex active contour and the Chan Vese model. Pattern Analysis and Applications. 2019;22: 343-357.

[13] Ali H, Lali MI, Nawaz MZ, Sharif M, Saleem BA. Symptom based automated detection of citrus diseases using color histogram and textural descriptors. Computers and Electronics in agriculture. 2017;138: 92-104.

[14] Zhang M, Liu X, O'neill M. Spectral discrimination of Phytophthora infestans infection on tomatoes based on principal component and cluster analyses. International Journal of Remote Sensing. 2002;23(6): 1095-1107.

[15] Füzy A, Kovács R, Cseresnyés I, Parádi I, Szili-Kovács T, Kelemen B, et. al. Selection of plant physiological parameters to detect stress effects in pot experiments using principal component analysis. Acta Physiologiae Plantarum. 2019;41: 1-10.

[16] Amato G, Falchi F. kNN based image classification relying on local feature similarity. Proceedings of the Third International Conference on Similarity Search and Applications. 2010;101-108.