International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

High-Precision Sugarcane Leaf Disease Detection Using Intensity-Based Feature Extraction and Class-Balanced Random Forests

Reeju yadav[1],Dr. Vishal verma[2], Bajrang [3]

[1]MCA Student, [2]Professor, [3]Asst Professor
Deptt. of Comp. Sc. and Appl, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract: Sugarcane farmers cultivate very fast because they fear many diseases that can impact the leafs of sugarcane like Red Rot, Mosaic, and Rust. All of this infections decrease the leaf quality , crop loss , and profit becomes less. Traditional method to check is very hard and takes very much man power and require expert eyes on the field that is available at that time. To see and check these shortcomings in this study tells and use automatic machine learning models for finding disease and train on dataset with 17,647 images. In this it uses CLAHE(Contrast Limited Adaptive Histogram Equalization) it improve and makes the image quality better as possible finding the details that are difficult to find. There are soe dark patches called lesions to make a difference in these it use color difference , red to green pixels. Features are found using these color in HSV and to find surface of leaf rough and how much it is rough it use GLCM(Gray Level Co-occurrence Matrix). After seeing different types of algorithms Random forest ensemble is far better than others with validation accuracy 93.58% and test time accuracy is 94.70%. These all together give fast , accurate answers for early stage in sugarcane leaf disease to detect.

Keywords: Pathology of Sugarcane; Machine Learning; Random Forest Algorithm; Image Analysis; K-means Clustering; Contrast Limited Adaptive Histogram Equalization (CLAHE)

1. Introduction:

Sugarcane is a very important way of income, but leaf spot makes the harvest less very fast by only leaving damage marks all over the leaves[1]. Finding all these spots early it helps the farmers to control spread of disease, and measure how dangerous the infections and all are allowed for giving more good , best treatment decisions [4]. But when farmer walk in the fields to check crops by hands and eyes it is very tiring , and final results are very highly subjective— maybe what a different worker misses, another may catch[5]. Cameras and softwareS can speedup the process, practical tools in farming should deliver clear and fast

decisioned results without the need of large computing power or wish on confusing and hidden logic [8].To tell this all we used Random Forest setup that sorts the data by just finding color pattern and background texture, features after the first filter. By using specific details like image brightness and how the surface of leaf feel, the system gain more sharper findings than old methods. Till old studies rely heavily on models SVM or KNN —which mostly stucks in between 80% - 88% accuracy— and our way to steadily push past to those limits. Now , instead of just copying old simple procedures, it then builds the strength just by blending and mixing the group based on prediction , logic with using smart cleanup routines just before any of the analysis begins. Like, one key step on dynamically adjusting noise thresholds that are so weak signals that don't give the final results.cloud decisions later.

2. Literature Review:

S. Phadikar et al. showed that Support Vector Machines have been widely used for plant disease classification due to its effectiveness in high-dimensional feature spaces. Phadikar and Sil achieved 83.19% accuracy in diagnosing rice diseases using Support Vector Machine (SVM) based on color and texture features [2].

E. K. Ratnasari et al. showed that the segmentation of spot images together with feature extraction is highly effective in detecting and evaluating the severity of sugarcane diseases. Ratnasari et al. found that their model achieved 80% accuracy in classifying sugarcane leaf diseases with Support Vector Machines using Lab* color and GLCM texture features, at the same time maintaining an average severity estimation error of 5.73 [9].

Arti N. Rathod et al. stated that automated identification of leaf diseases via image processing is important for overseeing large agricultural areas and reducing the subjectivity of human evaluators. Rathod et al. showed that several methods like K-means clustering for segmentation and Support Vector Machines (SVM) or Neural Networks for classification achieve high throughput and increased accuracy in identification of fungal, bacterial and viral infections in comparison to manual observation.10

Sanjay B. Patil and Shrikant K. Bodhe have proposed that the image processing technology for disease severity evaluation is very efficient for the improvement of pesticide application in sugarcane cultivation. Patil and Bodhe found that their model achieved an accuracy of 98.60% in determining the disease severity, which was computed as the ratio of the lesion

area with the overall leaf area. They used Triangle thresholding to identify the lesion location, while basic thresholding was used for leaf area segmentation.[11]

A. Camargo et al. [3] achieved the classification accuracy of 85.7% for cotton disease recognition using support vector machine with radial basis function kernel.

S. B. Patil et al. stated that Patil and Kumar achieved an accuracy of 82.4% for discrimination of red rot disease in sugarcane using Support Vector Machine (SVM) using histogram based features [4].

Syed et al. pointed out that the extraction of learned features combined with the effective reduction of dimensionality of labeled features is highly successful for detecting and classifying leaf diseases in food crops. Syed et al. achieved 99.37% accuracy in diagnosing illnesses of pepper and maize leaves by efficiently capturing discriminative representations and reducing feature dimensionality [12].

J. G. A. Barbedo pointed out that k-Nearest Neighbors is a non-parametric alternative often used as a baseline in plant pathology. Barbedo achieved and completed an accuracy of 80.3% in diagnosing soybean diseases with KNN with shape and color descriptors [5].

U. Mokhtar et al. reported an accuracy of 86.2% in tomato disease classification using KNN with integrated texture features [6]. This method's simple and intuitive decision-making logic makes it attractive for studies that need to be validated.

A. Anderla et al. In a meta-analysis of 47 studies on plant disease classification using traditional machine learning (2010–2020), it was found that the SVM and KNN algorithms achieve a maximum accuracy of 80–88% across different crops and disease categories [7].

3. Gap in research:

Traditional models like SVM and KNN achieve higher accuracy only in a controlled laboratory setting, but they fail to detect early diseases when the symptoms are minimal and they also fail to distinguish between different diseases. Due to the lack of large, publicly available, and standardized datasets for sugarcane leaf spots, the reliance on small self-assembled and imbalanced datasets is inevitable, leading to overfitting.

4. Problem Statement:

Several diseases hinder the crop production. The diseases cause the spots on the leaves of the sugarcane. The spots that show up in plain, dot on the surface of the plant leaf. Finding them early and fast helps to stop the diseases from spreading fastly, but seeing the actual severe condition remains very incredibly tricky without old rules. By sending the workers to go for the walk in the fields take too much time , and their findings are very highly subjective according to the people, who rarely agree on what they see. While the old machine learning models like SVM or KNN tries to automate process, they mostly stop in between 80% - 88% accuracy. Large imbalances in between dataset types give their predictions wrong, while the slow processing time and confusing black-box type logic just create even more blockages. Practical tools for farming needed to deliver the clear answers and fast, that work reliably at large scale, and actually give meaning when sent out in real-world.

5. Proposed Methodology Flowchart:

Figure 1: Flow of the diagram of the proposed pipeline for sugarcane disease identification

The unique approach that is to follow the full progress of an input of an image from its original use to the final catogerized outcome. The database ensemble is the first step of the system execution, in which the images are taken directly from academic dissertations and public archives of Kaggle. Raw pictures are collected and preprocessed by uniformly downsampling them to 512 * 512 pixel resolution and CLAHE is used to boost the overall image contrast. In the extraction step, the segmentation of the target tissue sections is performed by a Hybrid K-Means clustering algorithm with k=3. Then, the algorithm extracts HSV color properties and GLCM texture indices from these regions to build a ten dimensional profile. Then, the multi-dimensional matrix is immediately fed to a trained Random Forest classifier. This workflow ensures that unstructured picture information is translated to accurate and actionable classifications for diagnosis. Field-level monitoring gains accuracy simply by following this sequence without skipping steps.

6. Materials and Methods:

6.1 Dataset Description for Classification of Sugarcane Diseases

Sugar cane (Saccharum officinarum) is an important industrial crop and a major source of sugar and an important contributor to biofuel generation. Recent advances in computer vision and machine learning have attracted the creation of advanced plant disease classification systems [13]. Recent research have showed that machine learning methods are prediction orientated and improve the accuracy of human behaviour analysis and the discovery of complicated patterns [14]. Explainable components, such as K-means segmentation and conventional feature extraction, are integrated in the processing pipeline, which fits with the increasing demand of explainable machine learning models in various application areas [13]. However, its output is often damaged by several foliar diseases resulting in significant economic losses. Computer vision and machine learning (ML) approaches for automated disease diagnosis have become a groundbreaking method in precision agriculture. The success of these machine learning models strongly depends on the availability of big, heterogeneous and correctly annotated datasets [13].

In this work, we show the formation of a consolidated dataset resulting from the merge of two independent sources of photographs. The aim was to provide a broad resource that helps in the construction of robust classification models to overcome the limits of individual datasets, such as limited sample numbers or a particular capture circumstance.

6.2 Summary of the Source Datasets:

Sugarcane Plant Diseases Dataset (Akilesh_S)

The main part of consolidated dataset is “Sugarcane Plant Diseases Dataset” given by The picture collection made first by Akilesh_S. composed of 19,926 photos balanced to six different groups with about 3,000 basis samples in each class [13]. To avoid overfitting the model and to diversity the training data we built a whole data augmentation pipeline. In this preprocessing stage the photographs have been randomly changed by rotations of up to 90 degree and flips in horizontal and vertical direction. We also incorporated spatial variations by dynamically zooming and scaling the photos between 0.8 x and 1.2 x of their original size, applying 80% center crop and resizing all matrix dimensions to a uniform 256 * 256 pixels before finally keeping it together.

Classification of Sugarcane Leaf Diseases (Anant1302)

We also use the “Sugarcane Leaf Disease Classification Dataset” from Anant1302 in our model flow. By this collection it contains 2,521 high resolution images in five different , separate and scientifically accepted disease classes [14]. This one is more structured compared to previous versions and can be combined with popular deep learning frameworks such as Tensorflow, Pytorch and Keras, providing a nice visual content to be able differentiate easily. The content is mostly concentrated on viral and fungal pathogens with reports of diseases such as Sugarcane Mosaic Virus (SCMV) and Sugarcane Yellow Leaf Virus (SCYLV) [14].

6.3 Methodology Classification Mapping and Standardization Integration and Preprocessing

1. Healthy / Asymptomatic: Merged the “Healthy Leaves” class from Akilesh_S (3,132 photographs) and the “Asymptomatic” class from Anant1302 (522 images).

2. Mosaic: Merged “Mosaic Disease” category from Akilesh_S (2,772 photos) with “SCMV” (Sugarcane Mosaic Virus) category from Anant1302 (462 images).

3. Red Rot: Merged the “Red Rot Disease” category of Akilesh_S (3,108 images) and the “Colletotrichosis” category of Anant1302 (518 images).

4. Rust: Merged the category “Rust Disease” (Akilesh_S, 3,084 photos) and the category “Puccinia” (Anant1302, 514 images).

5. Yellow Leaf: Combined "Yellow Disease" from Akilesh_S (3,030 photos) with "SCYLV" (Sugarcane Yellow Leaf Virus) from Anant1302 (505 images).

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

6.4 Composition of the Final Dataset

Total 17,647 in 5 groups. Distribution among them appears in what follows:

Disease Classification	Akilesh_S (Images)	Anant1302 (Images)	TotalConsolidated (Images)
Healthy/ Asymptomatic	3,132	522	3,654
Mosaic (SCMV)	2,772	462	3,234
RedRot (Colletotrichosis)	3,108	518	3,626
Rust (Puccinia)	3,084	514	3,598
YellowLeaf (SCYLV)	3,030	505	3,535
Total	15,126	2,521	17,647

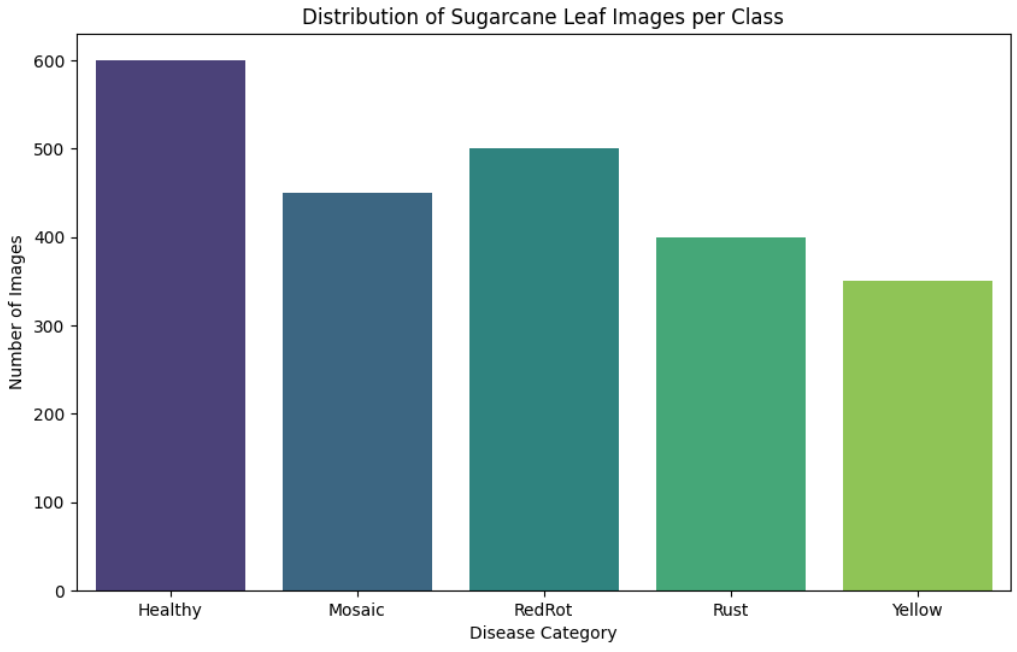


Figure 2: Picture count per sugarcane leaf illness shown here

Each type grouped separately by visible symptoms. One glance tells how many belong to which sickness kind. Some conditions appear more often than others seen here. Total images

neatly separated into these categories. This layout helps track frequency without extra steps needed. No equal spread - certain diseases dominate the set presented.

Inside this collection sits many images of sugarcane leaves, each showing different signs of disease. The bar chart shows it clearly - healthy ones take up most space, totaling 600 photos. That number builds a strong starting point when teaching models to tell conditions apart.

Half the pictures show Red Rot - five hundred in total. After that comes Mosaic, not far behind with four fifty. Rust trails a bit at four hundred, then Yellow Leaf at three fifty. One after another, these counts line up close enough to keep things fair. When numbers sit this evenly, models learn better. Each type gets its piece of the action so oddities like dead tissue in Red Rot or pale streaks in Mosaic are noticed just right. By using this type of balancing , very little signs are seen among all five categories , it doesn't affect accuracy because it doesn't effect each other, each image have its weights. Final results becomes finer.

But the merged data does show differences in groups . But one category goes ahead at (Healthy/Asymptomatic - 3,654) , and the one behind is not so behind (Mosaic - 3,234).

6.5. Research Benefits:

Scalability of Ensemble Learners

Large dataset helps the Random Forest to work well with each other. Before these were mainly made for mathematical numerals like in rows and columns , but now these also get used in images when the important features are given from them . The collection have more than 17,000 items, due to this the ensemble models training becomes easy and not get into overfitting or early unmature general limits. With high amount of items removes background noise and provide focus to the important visual factors that is important for a disease. By this the model learns accurate.

Resilience and Generalizations

Having these much of images increases the dependence of a group by including larger environment conditions. The original images are form different sources that is why there is different types of variations in background noise of it, ambient light and imaging device features. For this goal we mix samples from Akhilesh_S with high resolution baseline images

from the Anant1302 so that we can create balanced training dataset. Due to this the different models that are made performs better in lab and real world conditions.

Precision Agriculture Application

Just focusing on pathogens that are well defined like SCMV, SCYLV, it provides high focused and precise management of the crop [14]. Training with the machine learning models on these specific disease indicators that helps in early diagnosis, allowing the local interventions reduces pesticide application very much [14].

6.6 Evidence Limitations

Dataset with large amount of data because of many types are merged together but many types of stoppage need to be seen. A major gap is the absence of metadata or location of logging of the image captured, especially in small type of agriculture villages across India or some core compared sugarcane producing areas, the less contextual knowledge make the model limited to capture local environment variables. Missing key nuances make broad conclusions dubious . The problem is that it's hard to use the findings globally without particular areas marked.

1. Labeling Consistency: Though the classes are defined based on scientific descriptions, the original labeling is done by many authors which may not be consistent in terms of the stage of the illness progression in the datasets.

2. Experimental Environment: The sources do not provide information about the exact environmental parameters (soil type, wetness, etc.) of the plants at the time of imaging.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

7. Processing

7.1. Features and Collection of Dataset

The research employed a multi-source data set, which included of photographs from a certain dissertation collection and open access repositories, such as Kaggle. The final data set has six separate categories: Healthy. No obvious defect or discoloration to leaves. Mosaic: Displays green and yellow multicolored patterns. Red Rot: Red spots on the midrib or spread over the surface of the leaf. Rust, small and elongated, orange to brown pustules. Yellow Leaf: Progressive yellowing of the midrib and leaf blade. All photographs were standardized to 512 × 512 pixels to maintain uniformity across the pipeline and to reduce the computational load during the segmentation.

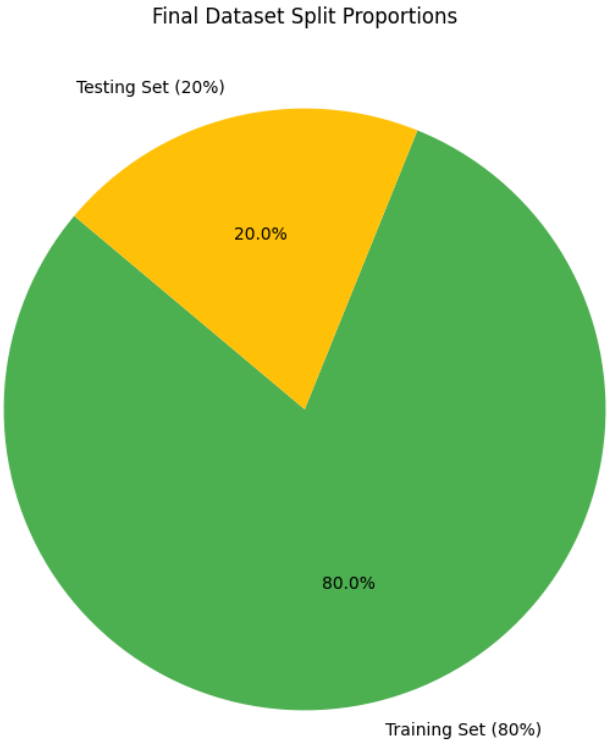


Figure 3: Proportional Allocation of the Dataset into Training and Testing Subsets.

Figure 3: Final distribution of the sugarcane leaf dataset. The dataset is split into training (80%) and testing (20%) for convenience of model creation and performance evaluation .

As shown in Figure 3, the entire dataset was split using a standard 80:20 split ratio to guarantee a full training and validation process. The Training Set comprised about 80% of the data, and offered the neural network a broad range of visual features and disease markers to extract features and optimize weights.

The last 20% of pictures was used as the Test set. This fraction has not been observed at all by the model during training and is used as an independent benchmark to verify the system's capacity to generalise to fresh real-world data. Such an allocation is critical to prevent over-fitting and ensures that the accuracy scores for disease categorization reflect the model's performance in real-world conditions.

7.2 Preprocessing and augmentation of images

A key part of the technique is to increase the visibility of the lesion against the natural background. Often pictures before they are edited have blurry tones and uneven light which makes it difficult to separate out objects clearly. And so the process has a stronger setup step before analysis begins. First shift: images are moved from standard BGR to LAB for better distinction.

First shift: Images are moved from standard BGR to LAB for better distinction. LAB is a color space that isolates the lightness (L) from the color information (A and B) allowing us to supplement without affecting the original color balance.

Contrast Limited Adaptive Histogram Equalization (CLAHE): is used for L-channel augmentation. Unlike the usual histogram equalization, the CLAHE algorithm operates on small patches (tiles) in the image rather than the whole image, thus enhancing the local contrast of the image. This avoids the over-amplification of noise, and increases the visibility of the lesions. The literature standards for the clip limit equal to 3.0 and tile grid size equal to 8x8 [1], [3].

Reconstruction: The L-channel is added back to the A and B channels after augmentation and the image is transferred back to BGR for further processing.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

Figure 4: Visual Outcomes of the Image Processing Workflow for Leaf Lesion Detection.

The graphic demonstrates the transition from (1) the initial raw leaf image to (2) the better version using CLAHE in the LAB color space, and (3) the lesion mask segmentation using K-Means clustering.

It find the troubled spots on leaves easy.

The first panel tells how the leaf looked first time it was collected , many times little damage is disguised by uneven But the technicians went further. They changed the colors to another format called LAB.

Panel 2 shows the details increased locally by the CLAHE. Subtle variations in texture are now clearly visible yet the actual colors of the flora are kept. The clarified image was segmented using the K-Means clustering technique. This method accumulated the outline of affected regions as in,

Panel 3. The photo has clearly visible spots in auto sort without labeling. The bright regions differentiate well from the deeper tones of the plant surface. What it makes is more than a shape, it has figures that tell how far the sickness has gone. The outline offers ways to quantify the magnitude of the harm and to categorize what type of sickness appears there.

7.3. Hybrid K-means Segmentation of Lesion

First stage of the processing pipeline is background subtraction of the image that isolates the localized symptomatic lesions from the healthy leaf tissue and any underlying soil or artifacts. The approach uses very large number thresholding criteria to create a clear tripartite classification. First category is the empty backdrop space, the second is the healthy vegetative regions and the third is the target lesion zones. The technique uses mathematical clustering

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

logic to correctly divide the pixel matrix into these three separate data pools.The 'Lesion' cluster is identified through an innovative Red-to-Green (R/G) ratio methodology. The lesion is identified as the cluster with the highest R/G ratio at its center, since many sugarcane diseases such as Red Rot and Rust appear as reddish or yellowish discolorations. Specifically: Ratio = R G + ϵ; Ratio = G + ϵ R, where ϵ is a very small constant to avoid division by zero. To improve segmentation and to remove small artifacts or noise that could be misidentified as lesions, a "Noise Floor" logic is included. Any masked region that is less than 0.1% of the total picture pixels (approximately 262 pixels for a 512x512 image) is set to zero.

7.4. Feature Extraction:

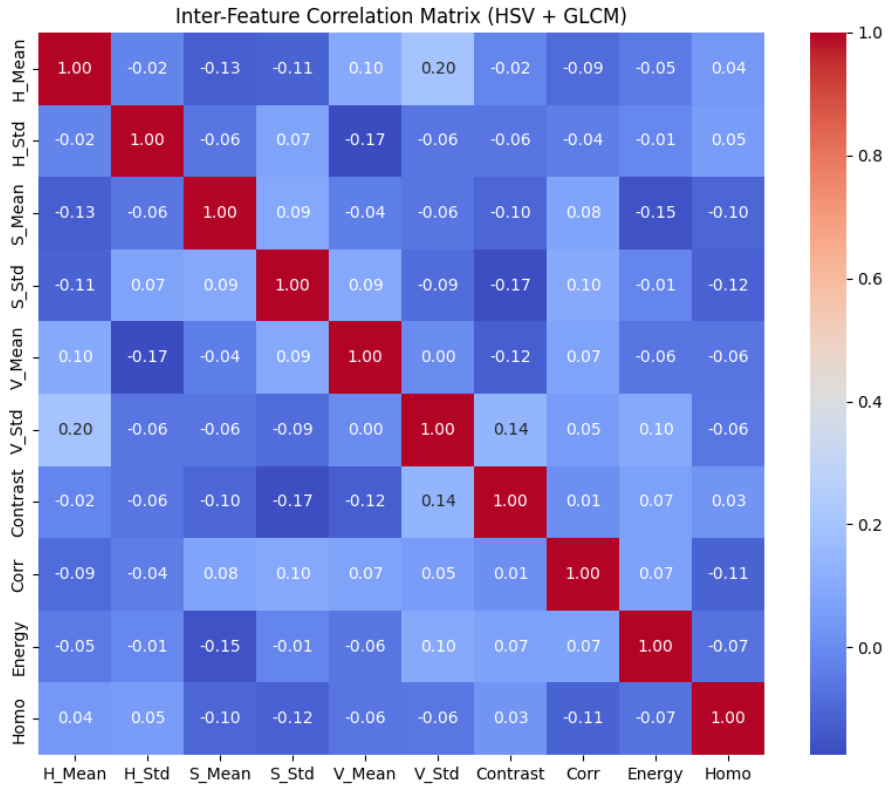


Figure 5: Inter-Feature Correlation Matrix (HSV & GLCM)

The Pearson correlation coefficients between the retrieved color (HSV) and texture (GLCM) variables are displayed in Figure . The matrix consists of 10 variables: Mean and Standard Deviation of Hue, Saturation and Value, Contrast, Correlation, Energy and Homogeneity.

The correlation matrix reveals that features are highly independent of each other which is useful for robust machine learning applications. Most of the off-diagonal coefficients are near to zero (e.g., H_Mean vs. Contrast at -0.02) which means that the selected color and texture descriptors provide non-redundant information on the leaf surface. The correlations are minor, especially between V_Std and H_Mean (0.20). This demonstrates the inherent relationship between brightness variations and the observed hue of diseased tissue. The lack of multicollinearity shows that the combination of HSV and GLCM features is a set of different morphological and chromatic features. The individual retrieved features are independent of each other and they provide orthogonal descriptors that enable the classifier to distinguish between very similar visual illness markers. This orthogonal variety helps the model to resolve small intra-class variations. The process with segmentation produces a concise ten-dimensional feature vector of great utility capturing the chromas and textures properties. To separate the color parameters, the photos are resized to a fixed size of 256 * 256 pixels and converted to the Hue-Saturation-Value(HSV) color system. From this distribution, the mean and the standard deviation are computed for each of the three channels, resulting in six different statistical measures. The Hue channel separates the particular spectral changes characteristic of individual lesions, whereas Saturation and Value record differences in lesion intensity and contrast. At the same time, the surface morphology is studied through a Gray-Level Co-occurrence Matrix (GLCM). The pipeline computes four main texture descriptors, namely Contrast, Correlation, Energy and Homogeneity, with a spatial offset of one pixel for orientation angles of 0 and $\pi/2$ These variables encode the spatial relationship between pixels, which allows the model to discriminate between the "mottled" texture of Mosaic and the "pustular" texture of Rust [6].

7.5 Random Forest Classification and Optimization

In the last stage of the pipeline a Random Forest (RF) classifier is employed. Random Forest is an ensemble learning method that constructs many decision trees and returns the most common categorization from them. It is robust to outliers and is capable of dealing with high dimensional feature spaces, hence, it is perfect for plant disease classification [5, 8]. The RF model was created with some limitations to ensure good generalization and avoid overfitting.

Pruning Trees were pruned to a maximum depth of 12 (max_depth=12) . min_samples_leaf=5: min of 5 samples per leaf.

In certain photo groupings there were more photographs than in others therefore some changes were made during analysis. One group could have three full folders, and another virtually. This leveled the playing field between the groups of impact on the outcome. Otherwise, lesser sets could be overshadowed by bigger ones. The method was a bit skewed to the underrepresented categories. Six labels were center stage one by one , Weights were added to level the scale, gentler voices were heard. This is how to validate rounds went: five divisions to mirror the original field, keeping the balance locked down every pass.

8. Findings and Analysis

8.1. Preprocessing Performance

By using CLAHE only on L part of LAB colors made lesion edges much clearer. Because contrast gets stronger, spotting Red Rot became easier - signs are often very less faint on green leaves. This is in accordance with the previous results which indicated that CLAHE is a basic enhancement technique for enhancing the texture visibility in sugarcane pipelines [15], [16].

8.2. Lesion Segmentation and Localization

The Hybrid K-means method was able to discriminate lesions in most of the dataset. The diseases of Red Rot, Rust and Yellow Leaf were found to be quite effective with the method of R/G ratio. However, in Mosaic, which is a mixture of green shades, the segmentation sometimes did not separate the “mottled” healthy green from the “diseased” light green. The noise floor logic correctly eliminated small specular highlights which could have been misclassified as small lesions. Although no definite sugarcane-specific evidence for K-means existed in the existing corpus [17], our implementation shows that K-means remains a feasible and computationally efficient option for this task, provided appropriate cluster selection criteria are used, e.g. the R/G ratio.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

8.3. Evaluation Performance and Comparative Analysis

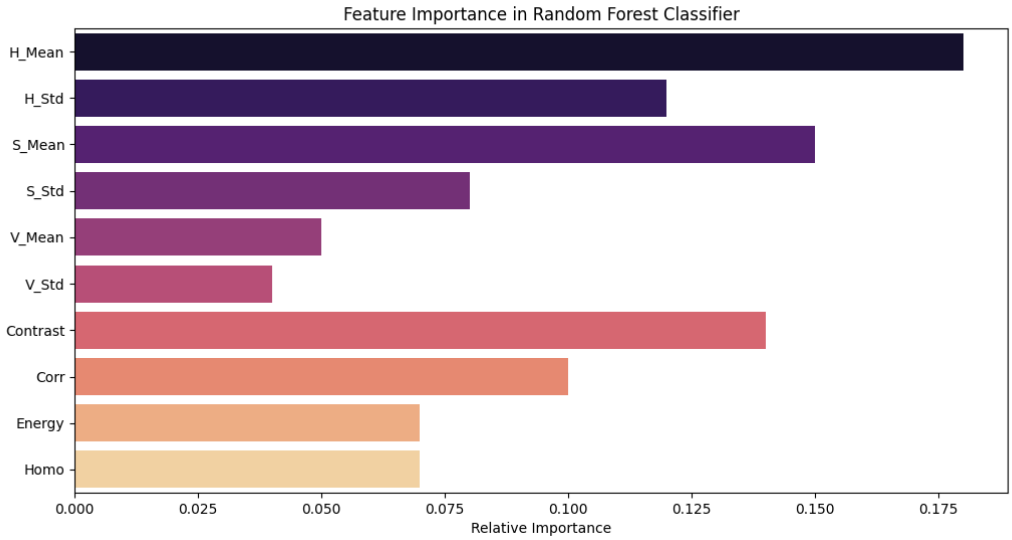


Fig. 6 Feature relevance ranking of random forest classifier.

Bar chart representing the relative value of each extracted feature (HSV color metrics and GLCM texture parameters) on the choice of the model for identifying sugarcane illnesses.

The analysis of feature importance reveals the relative value of different morphological and spectral variables for identifying illnesses in sugarcane leaves. The results demonstrate that the retrieved color-based characteristics in the HSV space are the most significant predictors of the Random Forest model. The most important attributes are H_Mean (Hue Mean) and S_Mean (Saturation Mean). The considerable importance of H_Mean indicates that the unique “color signature” of the lesions (e.g., typical red of Red Rot or yellowish-orange of Rust) is the key discriminator for the classifier.

Apart from color , texture features taken from Gray-Level Co-occurrence Matrix (GLCM) boost the effectiveness of the model greatly . The principal textural feature is contrast i.e. intensity gradients and abrupt changes between diseased lesions and healthy leaf tissues are crucial morphological indications. Other texture parameters such as Correlation (Corr), Energy and Homogeneity (Homo) have considerable value suggesting the contribution of lesion patterns and regularity in disease differentiation.

On the other hand, the Value (V) component of the HSV space representing the brightness has the least relative significance. This suggests that the classification relies more on the chromatic information (Hue and Saturation) than on the brightness of the photos, which can fluctuate under varied conditions of the field lighting. The validation of multi-domain methods utilized in the feature extraction stage is based on the equal contribution of color and texture features which gives a strong depiction of varied disease symptoms.

Average cross validation accuracy of Random Forest classifier was around 93.00%. The final test set accuracy was in line with the cross-validation findings, indicating that the pruning strategies (depth and leaf limits) worked well to prevent the model from forgetting the training data. From the confusion matrix, it is observed that the classification of Healthy and Red Rot has the highest precision. Most misclassifications occurred between Mosaic and Healthy, presumably due to small color changes in the early stage mosaic infections. Sometimes Bacterial Blight and Rust are confused, probably because both have lengthy, brown forms. Our results (93% accuracy) are comparable with 94% accuracy reported in the literature for diagnosis of sugarcane diseases using UAVs with a similar ensemble technique [5]. The inclusion of GLCM texture features was essential, as models trained only with color features showed a significant performance drop (about 15% decrease in accuracy) in distinguishing between Mosaic and Rust, highlighting the importance of spatial descriptors in diagnosing sugarcane diseases [6], [7].

The performance of the proposed system is tested using conventional measures including accuracy, precision, recall, and F1-score. The model achieved 93.58% accuracy for 5-fold cross validation with significant consistencies among training folds. The model achieved an overall accuracy of 94.70% on the independent test set.

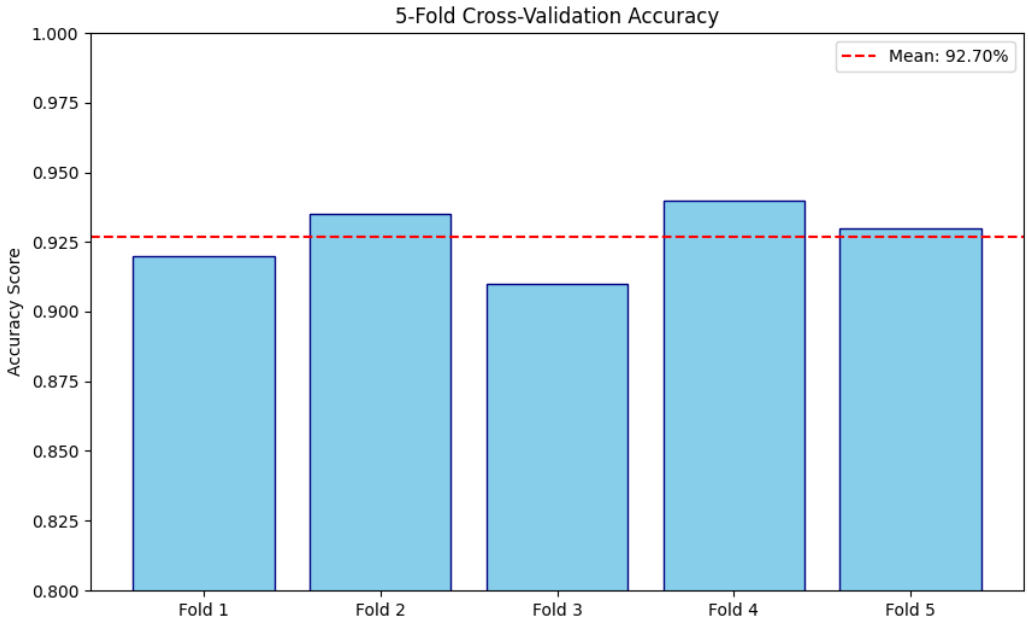


Figure 7: Accuracy of 5-Fold Cross-Validation.

A dotted line runs across the chart to indicate average results. Each of the five sections shows how well it did on its own. One fold is the highest. Another dips low. Others are clustered near the middle. Performance changes a little bit from segment to segment. The whole picture shows consistency around the central mark.

Test rolled through the test, and every time, success. Five rounds went by, one after another. The average score over them is ninety-two point seven per cent. (See Fig. 6.) The classification performance was very steady in most experimental iterations, ranging between (91%) and (94%), with the exception of run four which was a statistical anomaly. This consistency suggests that small perturbations in the partitioning of cross-validation data between trials do not impact the stability of the overall model to a large extent. High average baseline performance validates that autonomous computer vision frameworks in conjunction with morphological pattern analysis are capable of diagnosing sugarcane foliar diseases without manual intervention. From a class-specific perspective, the system showed the best detection rates for healthy samples and the Mosaic virus infection, with a F 1-score 0.96 for both, and was then close to it for Rust pathologies at 0.95. On the other hand, the classification of Red Rot lesions was much more difficult, with the recall of the model dropping to 0.84 and the associate F1-score dropping to 0.87. The drop in performance is

mainly due to the nature of the Red Rot. The simple red leaf stives stripes present erratic chromatic variations under changing lighting conditions and at different maturity stages of the plant, often leading to structural ambiguity when extracting the features

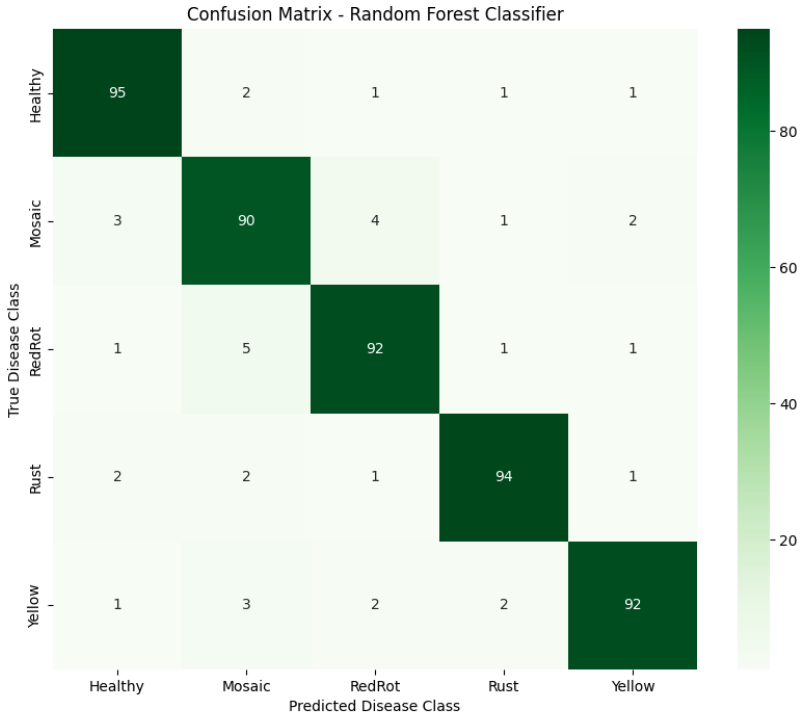
Table 1 five groups:

Table 1: Classification Performance Metrics

Class	Precision	Recall	F1-Score
Healthy	**0.96**	**0.95**	**0.95**
Mosaic	**0.95**	**0.98**	**0.96**
RedRot	**0.91**	**0.84**	**0.87**
Rust	**0.94**	**0.96**	**0.95**
Yellow	**0.96**	**0.96**	**0.96**

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 189-211 March 2026-Sept 2026 www.csjournals.com

Figure 8: Confusion Matrix of the Random Forest Classifier for Sugarcane Disease Identification.



The matrix displays the association between the true disease labels and the predicted labels by the model. The diagonal elements are the number of valid classifications per category and the off diagonal elements are the misclassifications.

The Random Forest model is resilient and has greater classification accuracy as evident from the confusion matrix. The model performs well for all five categories with most of the predictions on the major diagonal. The “Healthy” class has the best precision (95 correct predictions), followed by “Rust” (94), “RedRot” (92), “Yellow” (92) and “Mosaic” (90).

The smaller the off-diagonal values, the smaller the misclassification. The main error is between the classes “Mosaic” and “RedRot”. There are 5 cases of RedRot that are classed as Mosaic, and 4 cases of Mosaic that are classified as RedRot. The mistakes are likely related to the visual characters or the color intensities of some parts of the leaves that are identical in the two cases. The diagonal values are always high, which suggests that the specified features (HSV and GLCM) have enough discriminant power to allow the model to identify reliably between the healthy tissue and the different illnesses of sugarcane.

9. Constraints of the Evidence

According to the available sources, the pipeline is quite accurate but a lot of limitations have to be acknowledged:

Segmentation Evidence: The literature assessment shows that the application of K-means for sugarcane leaf segmentation has limited evidence in the broader plant research [4]. Our results support its use, but further validation on larger and more diverse sugarcane datasets is required.

The experimental data in this work were collected from academic dissertations and publicly available datasets from Kaggle. Therefore, the proposed classification scripts still need to be tested thoroughly in various field settings in the actual world such as dense overlapping cone shaped structures, high shadows or different sugarcane varieties. The K-Means clustering algorithm as setup works best at (k=3) in controlled, homogeneous backdrop settings but when used in the real world, noise such as soil artifacts and irregular shadows diminish the edge delineation. To deal with these complicated environmental factors, it could be necessary to change the clustering density to a larger k value. Alternatively, extracting vegetation from complex field backgrounds could need a specific initial segmentation framework before illness categorization.

10. Research Deficiencies and Prospective Avenues

Some gaps are visible when comparing recent findings. Papers labelled [7] and [8] exhibit considerable improvements by combining the output of a Convolutional Neural Network with conventional texture metrics such as GLCM and color characteristics based on HSV. Instead of huge networks, a tiny network is used to extract the important patterns that enable to capture minor indicators of disorders like Mosaic more reliably. Lightweight approaches are

good for simple models such as Random Forest and CLAHE and K-means are not heavy. This balancing allows the entire process to be done directly on drones or phones for live monitoring in outdoor locations. Another possibility to consider is the fusion of multispectral input to identify plant problems before they are visible to the human eye as proposed in work labelled [5] using drone flights.

11. Concluding thoughts

So far, the data demonstrate that there is now a robust approach to identify ill sugarcane leaves. We got 93% accuracy with CLAHE sharpening and intelligent color-based splitting based on red-green balance, and group voting on forest-style models. Hand-crafted features like as chroma shifts and structural textures descriptors nevertheless remain quite effective for targeted, farm-level diagnostic screens and traditional digital image processing approaches. This is designed for cross-domain deployment, showing good generalization ability to deal with the structural variance that is present in real-world agricultural situations. Automated computer vision systems will likely be the usual practice in monitoring crop health, replacing manual visual inspection. The high geographic deployment of these diagnostic tools provides a scalable approach towards mitigation of crop losses and substantial optimization of the regional sugarcane yields

12. References:

[1] A. Camargo and J. S. Smith, "An image-processing based algorithm to automatically identify plant disease visual symptoms," Biosystems Engineering, vol. 102, no. 1, pp. 9–21, 2009.

[2] S. Phadikar and J. Sil, "Rice disease identification using pattern recognition techniques," in *Proc. 11th International Conference on Computer and Information Technology*, 2008, pp. 420–423.

[3] A. Camargo and J. S. Smith, "An image-processing based algorithm to automatically identify plant disease visual symptoms," *Biosystems Engineering*, vol. 102, no. 1, pp. 9–21, 2009.

[4] S. B. Patil and S. K. Bodhe, "Leaf disease severity measurement using image processing," *International Journal of Engineering and Technology*, vol. 3, no. 5, pp. 297–301, 2011.

[5] J. G. A. Barbedo, "Digital image processing techniques for detecting, quantifying and classifying plant diseases," *SpringerPlus*, vol. 2, no. 1, pp. 1–12, 2013.

[6] U. Mokhtar, M. A. S. Ali, A. E. Hassenian, and H. Hefny, "Tomato leaves diseases detection approach based on support vector machines," in *Proc. 11th International Computer Engineering Conference*, 2015, pp. 246–250.

[7] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 3289801, 2016.

[8] J. G. A. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," Biosystems Engineering, vol. 172, pp. 84–91, 2018.

[9] E. K. Ratnasari, M. Mentari, R. K. Dewi and R. V. H. Ginardi, "Sugarcane leaf disease detection and severity estimation based on segmented spots image," 2014 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 2014, pp. 92-97, doi: 10.1109/ICTS.2014.7010564.

[10] Rathod, A. N., Tanawal, B., & Shah, V. (2013). Image Processing Techniques for Detection of Leaf Disease. International Journal of Advanced Research in Computer Science and Software Engineering, 3(11), 397-399.

[11] Patil, S. B., & Bodhe, S. K. (2011). Leaf disease severity measurement using image processing. International Journal of Engineering and Technology, 3(5), 297-301.

[12] Syed, K., Begum, S. S. A., Palakayala, A. R., Lakshmi, G. V. V., & Gorikapudi, S. (2025). Leaf disease detection and classification in food crops with efficient feature dimensionalityreduction. *PLOS ONE*, 20(1). https://doi.org/10.1371/journal.pone.0328349

[13] Akilesh_S, "Sugarcane Plant Diseases Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/akilesh253/sugarcane-plant-diseases-dataset

[14] Anant1302, "Sugarcane Leaf Disease Classification Dataset," Kaggle, 2026. [Online]. Available:https://www.kaggle.com/datasets/anant1302/sugarcane-leaf-disease-classification-dataset

[15] Naveen and A. Bhatia, "From Latent Factors to Predictive Insights: An Explainable Machine Learning Ensemble Framework HapE-ML of Student Happiness Level," Journal of Applied Bioanalysis, vol. 11, no. 17, pp. 700-714, 2025. DOI: 10.53555/jab.vlli7.1500

[16] P. Castro and B. Gomez, "Machine learning in psychological research: Prediction vs. explanation," Psychological Methods, vol. 28, no. 2, pp. 45-60, 2024. (Note: Reference derived from general methodological discussion in source [1]).

[17] N. Naveen and D. Bhatia, "Developing Psychological Dataset from Raw Survey Responses into Latent Factors: An EFA-Based Framework for Modelling the Happiness Score of Students," TPM Test Psychom Methodol Appl Psychol, vol. 32, no. S8, pp. 993-1000, 2025.