ISSN : 2455-135X

Diagnosis of Disease for Sugarcane Crop Using AI

Rutuja Khose[1], Diksha Pawar[2], Pravin Dhole[3], Pravin Yannawar[4], Bharti Gawali[5]

[1] Dept. of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India
[2] Dept. of Computer Science and IT Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India
[3] Dept. of Computer Science and IT Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India
[4] Dept. of Computer Science and IT Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India
[5] Dept. of Computer Science and IT Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India

[1]rutujakhose5@gmail.com

Received: Revised: Accepted: Published:

Abstract - *The current study outlines an overall machine-learning model of the automated detection of disease-inducing factors in sugarcane foliage, using hybrid, image-based feature engineering coupled with more superior models of classification. A total of 2,521 images of five major disease groups, such as Healthy, Mosaic, Red Rot, Rust and Yellow Leaf Disease, were taken through preprocessing processes which involved resizing, denoising and translation into HSV colour space to enable increased chromatic languages. A hybrid feature-extraction approach, which involves the combination of classical feature descriptions HOG and deep convolutional features, which were achieved through EfficientNetB0, was used. These multidimensional hybrid features were further reduced to a 300-dimensional discriminative vector via the Principal Component Analysis hence significantly lowering the computational costs without losing any critical information. Several nearest classifiers, such as Random Forest and XGBoost were trained on the smaller set of features. The results of comparative performance analysis showed that XGBoost was more successful in performance as it reached a level of accuracy of 83.82 as opposed to the accuracy of 56.21 in the case of the Random Forest model. The consistency of the XGBoost classifier to separate disease categories that were similar was verified by further analysis with precision metrics, recall metrics, F1-score, and confusion matrices. The suggested system can be proven to be robust, efficient and scalable in detecting early disease in sugarcane crops.*

Keywords - *Sugarcane Disease Detection, CNN, XGBoost, PCA, Disease Prediction*

1. Introduction:

Sugarcane forms a crucial sector in the agrarian economy of India and is a source of immense contribution to the agricultural industry and the livelihood of the rural population [1]. Since the country is second in production of sugarcane in the world, the crop supports the lives of millions of farmers involved in the production of sugar, jaggery, ethanol and biofuels [2]. However, sugarcane also is exposed to a range of foliar diseases such as red rot, smut, mosaic and rust, which causes significant losses in yields unless timely interventions are taken. Traditional methods of disease detection are highly dependent on subjective expertise, manual inspections and laboratory diagnostic methods, making them time-consuming, subjective, unfeasible in large agricultural settings. With the introduction of computational methods of image-processing and machine-learning, such alternative methods of disease identification have become very strong. Most of the analytical techniques that have been widely used to obtain discriminative features in leaf images in recent academic studies include colour histogram, Haralick texture descriptors, Local Binary Patterns (LBP), and Gabor filtering. Classification algorithms such as random forest and XGBoost have been found to have high precision of diagnosing plant diseases [3].

Dimensionality reduction methods like Principal Component Analysis (PCA), feature selection algorithms and variant filter methods have also taken further measures to improve model performance by identifying the most relevant features.The current work is devoted to the creation of the effective machine-learning architecture in the field of automated sugarcane leaf disease classification [4]. Data was used consisting of 2,521 images which consists of five major categories including: Mosaic (462 images), Red Rot (518 images), Rust (514 images), Yellow Leaf Disease (505 images), and Healthy leaves (522 images). In order to obtain uniform, high-quality inputs, every picture was subjected to pre-processing in the form of resizing to 128x128pixels, removal of noise through the use of the Gaussian blur filter, and HSV colour space conversion that is better at reflecting disease-related changes in colour as compared to the RGB format.

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

2. Literature:

Prior to the year 2020 the methods of detection of the disease of the sugarcane were founded on the classical approaches to the image-processing and visual inspection. These techniques offered handcrafted properties like color, texture and shape descriptors and machine-learned properties, including Support Vector Machines (SVM), Random Forest, and K-Means. Despite the fact that, the methods led to automation, these methods were subjective and dependent on the environmental factors and could not be applied on the level of the complex images of the real field [1].

Deep-learning-based (especially Convolutional Neural Networks or CNNs) studies became the trend in the year 2020. They were discovered to possess the ability of extracting profound features of pictures of leaf spontaneously hence, do not require person-by-hand feature extraction [6]. The comparative works showed that the CNN-based transfer learning models are more accurate and robust compared to the antiquated machine-learn techniques. The image- processing and manual feature extraction were not researched to the same scale but it was obvious that deep learning was superior [7].

In 2022, the researchers extended the application of such techniques to multispectral images taking place by the UAV and transfer learning to identify the diseases of sugarcane, including white leaf disease. Through these works, the higher importance of aerial survey imageries and precision farming technologies in the mass disease surveillance was given [10].

In 2023, a number of studies had determined that manual inspection is a time-consuming, expensive and cannot be used in a large agricultural farm. The most popular CNN networks in detecting the disease of plants are VGGNet, AlexNet, GoogleNet, ResNet and DenseNet, which have been demonstrated to achieve a constant higher accuracy than the traditional machine- learning models. The same year also has other works that highlighted the significance of both heterogeneous datasets and effective feature-extraction models to enhance generalization of the models [9].

As early as 2023, new methods resulted in deep neural networks of spectral-spatial attention- based hyperspectral imaging. The sugarcane diseases such as smut and mosaic were quite simple to identify at an early stage before the diseases progressed as a result of these models. The excellent hyperspectral data and attention-based features learning yield high accuracy, robustness and sensitivity over the traditional CNNs, which makes such systems applicable in scalable high-precision agriculture systems [13].

In 2024, comparative studies in large scale, which compared various pretrained CNN models, including VGG19, ResNet18, AlexNet, DenseNet201, ResNet101, DenseNet161 and MobileNetV2, were used to classify sugarcane leaf disease. It was continually determined that version of DenseNet and ResNet were the most effective ones normally with an accuracy of over 97 percent and with reduced feature reuse and gradient drop out [5].

The study on how ensemble and hybrid deep-learning models may be trained was also done i.e. it used several CNN models (e.g., VGG, ResNet, DenseNet) to enhance extrapolation to novel environmental conditions i.e. in the year 2024. These group techniques were more consistent and more true in asserting the various datasets [11].

In 2024, other researchers compared deep-learning methods with the traditional machine- learning models that employed chromatic, textural and morphological features. The SVM and Rand Forest were also good concerning the fact that they are used as the baseline model, but CNN as a transfer learning was superior at all times. The integration of the UAV multispectral imaging and CNN in detecting diseases was also an enhancement of the capabilities of the disease detection in the natural field setting [4].

All the review papers were done in 2024 to determine the effectiveness of both machine-learning methods and deep-learning methods in detecting crop diseases. According to such reviews, there was a challenge as regards the percentage of classes particularly regarding the application of famous datasets such as the PlantVillage that harms the generalisation of the models. However, CNNs that were trained with VGG16, ResNet50 and DenseNet121 provided a recognition rate of 95-99 percent because they could learn complex spatial hierarchy [3].

It was developed in order to make some progress by 2025 with studies on deployable and explainable systems. The sugarcane disease was detected in real time by the deep-learning structures that were equipped with deep-resolution imageries, IoT devices, drones, and smart cameras. The explainable AI models including Grad-CAM were used to enhance the model trust and openness. It recorded 97.2 percent accuracy using the systems but this showed that the systems can be applied in the actual agriculture environment [2].

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

3. Methodology:

The methodology involves the process of acquiring sugarcane leaf image, preprocessing of the acquired leaf images, extracting both the classical and deep-learning derived features, a dimensionality reduction of the features set and training machine-learning classifiers in disease recognition [3]. After the model is trained, its effectiveness is carefully tested and the completed model is made to predict the disease conditions using new leaf images.

3.1. Dataset:

The sugarcane disease sample data from Kaggle used in this research study which has 2,521 leaf pictures, and all have five major classes which are as follows: Mosaic (462 images), Red Rot (518 images), Rust (514 images), Yellow Leaf Disease (505 images), and Healthy leaves (522 images) [14]. There is even distribution of the samples among the classes, thus a sufficient number of observations are obtained in creating a valid deep-learning structure with the objective of classifying sugarcane disease. Image processing was done to normalize input image and improve the performance of features that were obtained in the case of leaf images. All the images were read out firstly in the dataset and then resized to 128 x 128 pixels, thus maintaining the same dimensions and minimizing the complexity of calculation. In order to reduce noise and accentuate disease patterns, a Gaussian blur filter was used. After the denoising, the image was encoded using an HSV (Hue Saturation Value) colour space by transforming the RGB colour space, and this is better to encode the colour variations related to disease symptoms. The hue and saturation features help to extract the features more effectively as the hue and saturation represent the finer colour variations amidst diseased and healthy leaf areas. Together, this preprocessing pipeline generates more standardized and cleaner images with colour enhancing features that are more easily extracted and classified.

3.2. Preprocessing:

The sugarcane leaf raw images are pre-processed through a pipeline that focuses on creating uniformity and reducing noisy features before features are extracted. OpenCV is used to read each image and resize it to 128 X 128 pixels, therefore, converging an operative input size to the same size throughout the dataset and minimizing later computing power requirements [5]. In the current study every leaf image is first optimized to boost visual quality and make feature extraction to be more reliable. The image is then transformed to the Hue Saturation Value (HSV) colour model which is less prone to changes in illumination as compared to the RGB. It is ensured that there is histogram normalization to ensure that there is similarity in image scale with dissimilar images. Overall, any images are unsampled, denoised, and converted to some standard colour space (e.g. HSV/RGB) to enhance quality and accelerate accuracy of feature extraction.

3.3. Feature Extraction:

Classical descriptors are first of all extracted in a way that they describe elementary visual attributes of the leaf. The input image is greyed, and Haralick texture features are calculated using grey-level co-occurrence matrix (GLCM) method and this algorithm is implemented in mahotas library [4]. These texture features, smoothness, roughness, contrast, homogeneity and entropy are used to describe any salient leaf phenomena and help in the recognition of disease caused surface changes. An 8-time 8 binning scheme is then computed on the RGB channels to derive a three-dimensional colour histogram which the colour distribution patterns are then captured by, this is handy in that sugarcane diseases often occur in the way of colour variations, e.g. spots of yellow, red patches, mosaic patterns, or rust brown spots. Additionally, Histogram of Oriented Gradients (HOG) features are obtained to represent structure variations and leaf edge structures that make it easy to detect any vein abnormality, irregular boundaries, and morphologies of the lesions [7]. The texture, colour, and shape features formed by this are concatenated into one classical feature vector to produce a strong handcrafted feature of the leaf. At the same time, deep features obtained with the help of convolutional neural networks (CNN) are obtained with the pre-trained EfficientNet-B0 model. They are all resized to 224 X 224 pixels, normalised, and sent through EfficientNet -B0 (omitting the final layer), so enabling the network to learn high -level features on its own on the leaf images [8]. These CNN representations learn a subtle and complex set of patterns, including disease-specific textures, colour contrasts and faulty structure, that can be missed by classical approaches. Lastly, the classical feature vector and CNN embedding are joined together to build a hybrid feature one of which has a superior discriminative ability, stability, and effectiveness that reflect in significantly increased accuracy in sugarcane disease classification.

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

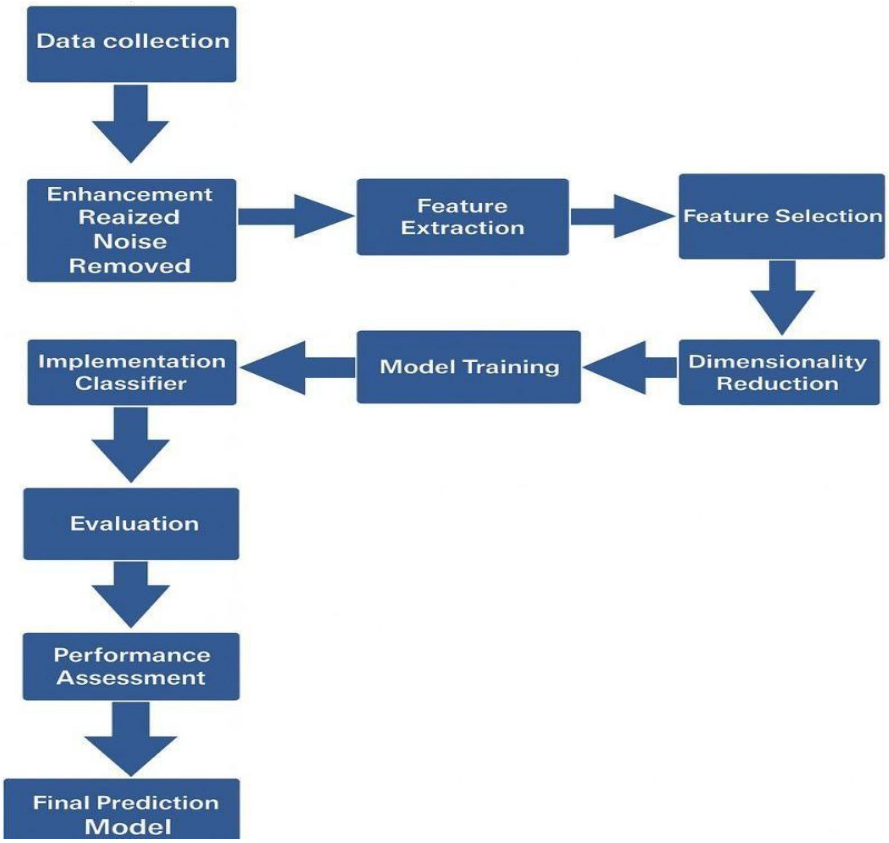

Figure 1:Sugarcane leaf disease prediction methodology

3.4. Feature Selection:

When it is synthesized, the resulting hybrid feature set has a very high dimensionality - it can easily contain many thousand values per image. This volume of data does not only increase computational times, but also increases memory utilization and predisposes the model to over fitting whereby noise is unintentionally trained as disease signatures. In order to reduce these issues, Principal Component Analysis (PCA) is used to reduce the size of the dimension. The unsupervised method (PCA) is a statistical method, which converts the original set of features into a smaller set of principal components each representing maximum variance of the data; PCA thus isolates automatically the most informative patterns, while eliminating redundancies and noisy features [7]. In this study, they used the PCA model to display the 300 top components which in a way reduces the hybrid feature set to a small but very informative ensemble. This dimensionality reduction drastically speeds up the training of the model, stabilizes performance of the classifier and improves generalisation by limiting overfitting [9]. After this, the trained PCA transformer is serialised to a.pkl file so that it can be used repeatedly in the process of inference.

3.5. Classifiers (Model Training):

The machine-learning classifier that is fed with the dimensionally reduced hybrid features is then used to classify sugarcane leaf diseases. First, the full set of features is divided into training and testing sets by an 80:20 stratified split thus maintaining a balanced distribution of classes [6]. Random Forest classifier is then trained using the training data; they are highly praised because of their robustness, ability to utilize high-dimensional data, and their ability to prevent overfitting which makes it appropriate when using the hybrid feature vectors. When the training is over, the model categorizes the test images into disease classes and the performance of the model is measured by means of standard measures, namely: accuracy, precision, recall, F1- score and a confusion matrix. Simultaneously, a XGBoost classifier is applied on the same reduced hybrid features. Once the model has been trained, the model call predict-proba gives back the probability scores of each disease type and argmax picks (Pick the disease with the highest predicted probability) the type with the highest probability as the ultimate prediction in each test example. The level of performance of the classifier is measured based on accuracy, a detailed classification report (including precision, recall, and F1-score of each class), and a confusion matrix. All these measures explain how much of the predictions were correct, the proportion of accurate and incorrect predictions in each category of diseases, as well as the distribution among the

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

falsely predicted groups and thus gives a complete picture of the diagnostic effectiveness of the model.

4. Result:

The overall accuracy of the XGBoost classifier on the test dataset of 507 images of sugarcane leaves was high at 83.82. In all five disease classes such as Healthy, Mosaic, Red Rot, Rust, and Yellow Leaf, the precision, recall, and F1-scores remained above 0.80 which implies that the predictions are consistent and reliable across the classes. It can be seen that the model correctly classifies most of the samples with 86% F1 with Rust and Yellow Leaf having the highest F1 with values of 90, showing that it can discriminate disease specific patterns despite the complex leaf textures. Comparing the XGBoost classifier with classical machine-learning models like Random Forest (accuracy 56.21) shows significantly improved and stronger results, which can be explained by its boosting learning algorithm, as well as its ability to better deal with hybrid representations of features. Assessment: The reduced hybrid feature set was used to train the Random Forest classifier and was assessed on 20% of the dataset that was used as a test set. The metrics obtained represent a total accuracy of 56.21, which is average classification ability. The precision, recall and F1-scores differed in accordance with disease classes and more success was achieved in certain classes (e.g., Healthy, Mosaic, and Rust) and poor results in others (e.g., Yellow Leaf and Red Rot). The confusion matrix indicates that the frequency of misclassification is rather common, which points to the fact that the Random Forest model is not proficient enough to uncover the complexities that abide in the images of sugarcane leaves disease. In general, the findings above demonstrate that the performance of the Random Forest is acceptable but inferior in comparison to more sophisticated classifiers like XGBoost or deep-learning solutions. The classification model was found to have an overall accuracy of 83.83 percent and it is very effective in the identification of several diseases in sugarcane. When it comes to per disease per class, there are good perceptions and recollection with Yellow and Rust disease and F1-score up to 0.90. Competitive metrics are also exhibited by Mosaic and Red Rot to show balanced prediction ability in all the categories. The confusion matrix also verifies a consistent and sound classification behavior with majority of the samples placed in their infection classes correctly.

Table 1:Classification report of Random Forest

Random Forest				
No	**Precision**	**Recall**	**F1-score**	**Support**
Healthy	**0.54**	**0.63**	**0.58**	**105**
Mosaic	**0.70**	**0.45**	**0.55**	**94**
Red Rot	**0.52**	**0.61**	**0.56**	**104**
Rust	**0.66**	**0.63**	**0.65**	**103**
Yellow	**0.47**	**0.49**	**0.48**	**101**
Accuracy: 56 %				

Table 2: Classification report of XGBoost

XG Boost				
No	**Precision**	**Recall**	**F1-score**	**Support**
0	**0.79**	**0.90**	**0.84**	**105**
1	**0.86**	**0.76**	**0.80**	**94**
2	**0.78**	**0.81**	**0.79**	**104**
3	**0.89**	**0.83**	**0.86**	**103**
4	**0.90**	**0.89**	**0.90**	**101**
Accuracy: 84 %				

5. Conclusion:

The paper presents a complete and systematic system of the automated detection of sugarcane diseases through machine learning using images. To test and train the proposed system, a collection of 2,521 leaf images representing five main types of diseases, including Mosaic, Red Rot, Rust, Yellow Leaf Disease, and Healthy, was used. The process will involve necessary steps such as image preprocessing, classical features creation and convolutional neural network (CNN) with deep features generation, hybrid feature fusion, dimensionality reduction via principal components analysis (PCA) and machine-learning-based classification. The classical module of feature obtained discriminative colour, texture and structural features. These complementary features were summed together to create a hybrid feature and PCA was used to reduce the size of the hybrid feature to 300 dimensions, keeping computation methods small and curtailing overfitting. Random Forest and XGBoost were tested as two classifiers using reduced feature set. The Random Forest was only able to obtain a moderate accuracy of 56.21%, which indicates that the problem of characterizing

ISSN : 2455-135X \

the complex feature space is difficult. Conversely, XGBoost achieved an accuracy of 83.82% and it has a better ability to model non-linearized relationship and to model high-dimensional integrated features. The incidence of improvement of the precision, recall, and F1-score support the claim that XGBoost is a more robust and dependable classifier of sugarcane leaf disease.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information

This research did not receive any specific grant from

funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

The authors express their gratitude towards Dr. Babasaheb Ambedkar Marathwada University in Maharashtra, India, provided valuable support.

ISSN : 2455-135X \

References:

[1] Reddy, A. V., Thiruvengatanadhan, R., Srinivas, M., & Dhanalakshmi, P. (2023). Artificia l intelligence framework for sugarcane diseases classification using convolutiona l neural network. *Int. J. Recent Innov. Trends Comput. Commun*, *11*(9), 3620-3628.

[2] Akhil, S., & Durga, R. (2025, June). Analysis and Detection of Diseases in Leaf Images for Sugarcane using Efficient PARNET-52 Deep Learning Techniques. In The 2025 International Conference on Advanced Research in Electronics and Communication Systems (ICARECS-2025) (pp. 121-133). Atlantis Press.

[3] Vivekreddy, A., Thiruvengatanadhan, R., Srinivas, M., & Dhanalakshmi, P. (2024). Artificial Intelligence Framework for Multi-Class Sugarcane Leaf Diseases Classification Using Deep Learning Algorithms. Journal of Theoretical and Applied Information Technology, 31(10).

[4] Bao, D., Zhou, J., Bhuiyan, S. A., Adhikari, P., Tuxworth, G., Ford, R., & Gao, Y. (2024). Early detection of sugarcane smut and mosaic diseases via hyperspectral imaging and spectral-spatial attention deep neural networks. Journal of Agriculture and Food Research, 18, 101369.

[5] Nomani, K. M. S., Nuruzzaman, M., Nadia, A., & Billal, M. M. (2024, October). Sugarcane Leaf Disease Detection: A Comparative Analysis Using Deep Learning. In Proceedings of the 3rd International Conference on Computing Advancements (pp. 139-144).

[6] Islam, M. M., Adil, M. A. A., Talukder, M. A., Ahamed, M. K. U., Uddin, M. A., Hasan, M. K., ... & Debnath, S. K. (2023). Deep Crop: Deep learning-based crop disease prediction with web application. Journal of Agriculture and Food Research, 14, 100764.

[7] Nomani, K. M. S., Nuruzzaman, M., Nadia, A., & Billal, M. M. (2024, October). Sugarcane Leaf Disease Detection: A Comparative Analysis Using Deep Learning. In Proceedings of the 3rd International Conference on Computing Advancements (pp. 139-144).

[8] Wadhe, V., Dongre, R., Kankriya, Y., & Kuckian, A. (2022, December). Sugarcane disease detection using deep learning. In 2022 5th International Conference on Advances in Science and Technology (ICAST) (pp. 40-44). IEEE.

[9] Upadhye, S. A., Dhanvijay, M. R., & Patil, S. M. (2022, December). Sugarcane disease detection using CNN-deep learning method. In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (pp. 1-8). IEEE.

[10] Santhrupth, B. C., & Devaraj Verma, C. (2024). Intelligent disease detection in sugarcane plants: A comparative analysis of machine learning models for classification and diagnosis. International Journal of Intelligent Systems and Applications in Engineering, 12(8s), 299-306.

[11] Ngugi, H.N., Akinyelu, A. A., & Ezugwu, A. E. (2024). Machine learning and deep learning for crop disease diagnosis: Performance analysis and review. Agronomy, 14(12), 3001.

[12] Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., ... & Ali, F. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science, 14, 1158933.

[13] Mathew, A. P., Viswajith, P., Rahman, A., & Murali, V. (2023). Crop disease detection using machine learning. Journal of Applied Science, Engineering, Technology and Management, 1(01), 28-32.

[14] https://www.kaggle.com/datasets/pritpal2873/sugarcane-leaf-disease-dataset