ISSN 2089-385X (Print) | 2829-6761 (Online), Volume 15, No. 1 (2025), pp. 1-6

Vertex

Published by: Institute of Computer Science (IOCS)

Enhancing Rice Disease Identification using Hybrid GLCM-XGBoost with SMOTE Imbalance Handling

Anwar Sadad

Health Informatics, Sunan Gresik University, Gresik, Indonesia

Abstract

Rice (Oryza sativa) is a major food staple, which is prone to multiple diseases that will dramatically decrease the harvest yield. Disease identification is time consuming and is usually subject to subjective errors in a manual approach. The following research will seek to increase the level of precision of automatic rice plant disease detection, namely the Brown Spot, Hispa, and Leaf Blast classes. The suggested method combines both the Gray Level Co-occurrence Matrix (GLCM) to extract texture features and the Extreme Gradient Boosting (XGBoost) classification algorithm. Furthermore, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to address class imbalance within the dataset of 5,548 images. Preprocessing steps include resizing, grayscale conversion, and Min-Max normalization. Experimental results demonstrate that the model trained on SMOTE-balanced data with optimized XGBoost parameters achieved a superior accuracy of 98%, outperforming the imbalanced scenario (97%) and previous studies. This research confirms that the combination of GLCM, SMOTE, and XGBoost constitutes a robust and high-precision method for rice disease identification.

Article Info

Article history:

Received : dec 05,2025
Revised : Dec 17, 2025
Accepted : Dec 30, 2025

Keywords:

Disease Classification;
GLCM;
Rice Disease;
SMOTE;
XGBoost;

Corresponding Author:

Anwar Sadad,
Health Informatics,
Sunan Gresik University,
Gresik, Indonesia.
anwarsadad@lecturer.usg.ac.id

Introduction

Rice (Oryza sativa) is a measured commodity in Indonesia. But the problem is that for achieving food security, Indonesia is increasingly faced with the threat of pests and diseases (Widyawati et al., 2025) The brown spot disease, leaf blast and hispa are included in diseases that are difficult to control, which greatly reduce the quantity and quality of the harvest. To overcome this, mobile surveillance and smart advisory solutions are increasingly being developed (Aida & Wan, 2025; Khalil et al., 2024; Talreja et al., 2022), owing to the need for accurate and early detection so that farmers can take the relevant control measures.

In recent years, the application of Machine Learning has proven highly effective in assisting precision agriculture and disease diagnosis (Nata et al., 2025; Priyangka & Kumara, 2021). A systematic review by (Seelwal et al., 2024) highlights the rapidly growing trend and potential of deep learning applications for rice disease diagnosis. Several studies have employed methods such as Convolutional Neural Networks (CNN) (Barburiceanu et al., 2021) hybrid Vision Transformer approaches (De Silva & Brown, 2023) and ensemble learning to automate detection. For instance,

(Sharma et al., 2021) highlighted the efficacy of deep learning in plant disease diagnosis, while (Tiwari et al., 2025) and (Kulkarni & Shastri, n.d.) successfully demonstrated the potential of multi-model machine learning for automated identification of rice diseases. Similarly, (Hutauruk, 2025) utilized YOLO algorithms for smart detection interfaces. Meanwhile, (Anggiratih et al., 2021) achieved an accuracy of 79.53% using EfficientNet B3, and (Wibisono & Saiful, 2025) demonstrated the effectiveness of XGBoost in handling agricultural datasets. Furthermore, hybrid approaches combining feature extractors with XGBoost have gained traction; for instance, (Sovia et al., 2025) successfully enhanced classification performance by integrating CNN with XGBoost.

Even though past findings have been promising, a significant research gap that should be filled is evident to enhance reliability. The major weakness of most of the existing studies is that they do not optimally deal with imbalanced datasets. (Miftahushudur et al., 2025) describe a recent survey that points out that one of the issues that need to be addressed to create powerful agricultural models is the imbalance in data. The imbalance in data may lead to the classification model being skewed on the majority class, thus leaving out the minority class that in most cases is the focus of the specific disease. Moreover, Deep learning is trendy, but hybrid techniques with statistic texture features are also very efficient and easy to compute (Alabbasi et al., 2025).

Conceptually, this study demonstrates that handling data imbalance in the feature space (via GLCM vectors) offers a more robust strategy for agricultural disease classification compared to traditional image-level augmentation. While the imbalanced model achieved 97% accuracy, it exhibited bias toward the majority class. The application of SMOTE on GLCM features successfully interpolated information gaps for minority classes, allowing the XGBoost algorithm to establish precise decision boundaries without the computational overhead of generating synthetic images (e.g., via GANs). This confirms that statistical feature-level oversampling is a highly effective, yet computationally efficient, strategy for enhancing diagnostic precision in smart farming applications.

While Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs), have shown remarkable performance in large-scale image classification, they are computationally intensive and notoriously data-hungry. Applying CNNs to medium-sized agricultural datasets often leads to overfitting unless massive data augmentation or transfer learning is employed. In contrast, rice leaf diseases—specifically Brown Spot, Hispa, and Blast—manifest primarily through distinct textural pathologies rather than complex shape variations. Consequently, statistical feature extraction methods like the Gray Level Co-occurrence Matrix (GLCM) can capture these fine-grained surface patterns more effectively than the latent features learned by CNNs in early layers.

Furthermore, a major limitation in existing literature is the lack of explicit control over class imbalance. End-to-end DL models often struggle with imbalanced data without complex loss function modifications. This study proposes a hybrid framework that decouples feature extraction from classification, allowing for precise intervention in the feature space using the Synthetic Minority Over-sampling Technique (SMOTE). By combining GLCM's texture descriptiveness with Extreme Gradient Boosting (XGBoost)—known for its robust regularization and execution speed —this approach offers a computationally efficient and highly accurate alternative to 'black-box' deep learning models.

To address these issues, the proposed study is a hybrid solution based on the idea of using Synthetic Minority Over-Sampling Technique (SMOTE) to deal with data imbalance, Gray Level Co-occurrence Matrix (GLCM) to extract texture features, and Extreme Gradient Boosting (XGBoost) as a classification method. GLCM is chosen because it is proven to be effective in the ability to capture finer texture leaf patterns verified by recent comparative analysis (Jordy & Ariatmanto, 2025; Ramli & Riadi, 2025). XGBoost is selected due to its high execution rate and regularization. The primary goal of the study is to enhance the accuracy of rice disease

identification by a considerable percentage over the previous methods in ensuring that the model is balanced in terms of learning.

Methods

Dataset The data used in the current paper has been gathered by the open Kaggle dataset repository ("Rice Leaf Dataset ") and single records. It has the total of 5,548 images grouped into four categories BrownSpot (1,138), Hispa (1,052), LeafBlast (1,561) as well as Healthy (1,797). The distribution of the classes in the dataset is very imbalanced with the majority of the classes being healthy as it is represented in the distribution.

Preprocessing To be able to transform the data in a uniform manner and make computations faster, the following data preprocessing actions were taken: Scaling and Grayscale Conversion: The images were scaled to 256x256 and converted to grayscale because so that emphasis was put on the texture variation and not on the variation in the colors. Normalization: Min-Max was used to bring the values to a range of between 0 and 1 in order to improve the speed at which the models converge.

Using Gray Level Co-occurrence Matrix (GLCM), we have obtained Feature Extraction (GLCM) Texture features. GLCM was determined at four orientations (0, 45, 90 and 135 0) and pixel distance d=1. GLCM can be applied in extracting features in the pathology of rice leaves because the current study has found that it is appropriate (Ramli & Riadi, 2025). Based on these matrices, there were six statistical values obtained; Contrast, Correlation, Angular Second Moment (ASM), Energy, Homogeneity and Dissimilarity that gave complete feature vector of each image.

Controlling Imbalance (SMOTE) To reduce the bias to the majority, Synthetic Minority Over-sampling Technique (SMOTE) was used on the training data. SMOTE provides synthetic samples of the minor classes (BrownSpot, Hispa and LeafBlast) by interpolating the existing samples. This was intended to equalize the dataset hence giving 1,797 samples of each class, and total of a total of 7,188 data points.

Classification and Testing Scenario This was divided into testing and training (80 and 20 respectively) set. XGBoost algorithm was used to do the classification. The hyperparameters were optimized through the grid search with: max depth, learning rate (eta), min child weight, n estimators and subsample and n estimators with the application of learning rate (eta). The Performance was measured with the help of the Confusion Matrix analysis, Accuracy, Precision, Recall and F1-Score.

Results And Discussion

Hyperparameter Tuning Results Hyperparameter tuning of the Extreme Gradient Boosting (XGBoost) model is very sensitive to hyperparameters. This research performed a set of experiments through Grid search to determine the best configuration. All 10 scenarios were experimented with different max depth, learning rate (eta), min child weight, n estimators and subsample.

The results of the parameter tuning are provided in Table 1. Experimental outcomes suggest that more profound trees (max_depth: 12) with an intermediate value of learning rate (eta: 0.3) and the increased number of estimators (500) were the most stable and most accurate.

Tabel 1. Results of Hyperparameter Tuning for XGBoost

Skenario	Tuning Parameter					Accuracy Balanced Data	Accuracy Imbalanced Data
	max_depth	Eta (learning_rate)	min_child_weight	n_estimators	Sub-sample		
1	6	0.3	1	100	1	97%	96%

2	12	0.3	5	500	0.9	97%	96%
3	12	0.3	1	100	1	98%	96%
4	6	0.3	5	100	1	97%	96%
5	6	0.3	1	500	1	98%	96%
6	6	0.3	1	100	0.9	98%	96%
7	6	0.3	5	500	0.9	98%	96%
8	12	0.3	1	500	0.9	98%	97%
9	12	0.3	5	100	0.9	98%	96%
10	12	0.3	5	500	1	98%	97%

Table 1 reveals that Scenario 8 had the best accuracy of 98% on the balanced dataset. The following parameters (max depth: 12, eta: 0.3, n estimators: 500) were then chosen as the best parameterisation.

The Effect of SMOTE on Model Performance A critical analysis was conducted to determine the effectiveness of SMOTE. Findings indicate that it is vital to deal with class imbalance. When the model was applied to the unbalanced data, it had a maximum accuracy of 97 with bias on the majority. The accuracy after using SMOTE was always 98. This is in line with the evidence presented by (Miftahushudur et al., 2025) that imbalance intervention is an effective way of enhancing the strength of agricultural classifiers.

Findings indicate that dealing with class imbalance is imperative. The model recorded a perfect accuracy of 97 on the unbalanced dataset. The bias was however found in that the model was more favourable to the majority group (Healthy). The accuracy was improved to 98 per cent regularly, after the use of SMOTE under the best conditions. SMOTE was able to create synthetic samples within the feature space of minority classes (BrownSpot and Hispa) which enabled the XGBoost algorithm to create a draw line of boundaries of decisions and minimize misclassification upon the disease categories.

Confusion Matrix Analysis In order to justify the thorough classification behavior, a Confusion Matrix analysis was applied to the most effective model (Scenario 8).

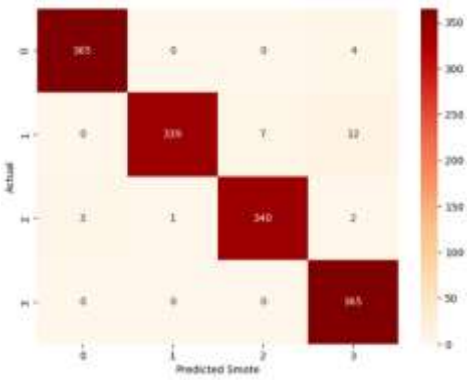

Figure 1. Heatmap Confusion Matrix

According to the analysis, the model is very precise. Particularly, the LeafBlast class was recognized with a high degree of fidelity and 0 false negatives were registered. Despite small error in the Healthy and Hispa classes, the error rate was very low. The mean Precision, Recall, and F1-Score of all classes were equal to 0.98 that proves the strength of the model.

Comparison to Previous Studies The proposed method (GLCM + SMOTE + XGBoost) was compared to the best methods provided in the recent literature. As can be seen as summarized in Table 2, the proposed approach performs better than a number of Deep Learning (CNN) and Machine Learning models.

Tabel 2. Comparison of Research Performance

No	Author	Method	Accuracy
1.	(Anggiratih et al., 2021)	Deep LearningEfficientnet B3 Dengan Transfer Learning	79%
2.	(Dubey & Choubey, 2024)	Feature Selection + Optimized XGBoost	86%
3.	(Kusanti et al., 2018)	Glcm & Backpropagation	80%
4.	(Khoiruddin & Tena, 2024)	Cnn	95%
5.	(Shrivastava & Pradhan, 2021)	Machine learning : SVM, DC, Knn, NB, DT, RF, LR	94%
6.	(Nata et al., 2025)	Cnn	91%
7.	(Milano et al., 2024)	Xgb	96%.
8.	**Research results**	**Xgb & Glcm**	**98%**

The result of 98% is better than most Deep Learning methods. This is an indication that although deep learning is strong, optimized hybrid feature extraction (GLCM) and ensemble learning can be very competitive and this is also echoed by (Alabbasi et al., 2025; Ramli & Riadi, 2025).

Conclusion

This study has demonstrated that the quality of agricultural disease identification is heavily dependent on how class imbalance is managed. The proposed method successfully integrates Gray Level Co-occurrence Matrix (GLCM) feature extraction with Synthetic Minority Over-sampling Technique (SMOTE) and Extreme Gradient Boosting (XGBoost), achieving a superior accuracy of 98%.

Conceptually, this research offers a critical update to data imbalance handling strategies in the smart agriculture domain by shifting the focus from image-level augmentation to feature-space interpolation. Unlike traditional methods that rely on geometric transformations or computationally intensive generative models, this study proves that applying SMOTE specifically on statistical texture descriptors (GLCM) effectively reconstructs decision boundaries for minority classes. This strategy eliminates the classification bias toward healthy plants without the computational overhead of image synthesis, offering a robust and lightweight solution. Future studies should consider hybridizing GLCM with deep learning embeddings or deploying this lightweight model into IoT-based monitoring systems as proposed by (Aida & Wan, 2025).

Reference

Aida, W., & Wan, N. (2025). *IoT Agri-Care Advisor Mobile Application for Monitoring Paddy Plant Health and Delivering Smart Farmer Advisory Toward Sustainable Agriculture Aplicación Móvil IoT Agri-Care para el monitoreo de la salud del cultivo de arroz y la entrega de asesoramiento inteligente al agricultor hacia una agricultura sostenible*. https://doi.org/10.56294/saludcyt20251979

Alabbasi, H. A., Abdulkarem, A., & Alrammahi, H. (2025). *Detection and Classification of Rice Plant Diseases Using Fusion Deep and Texture Features*. *14*(5). https://doi.org/10.18178/ijeetc.14.5.323-330

Anggiratih, E., Siswanti, S., Octaviani, S. K., & Sari, A. (2021). Klasifikasi Penyakit Tanaman Padi Menggunakan Model Deep Learning Efficientnet B3 dengan Transfer Learning. *Jurnal Ilmiah SINUS*, *19*(1), 75. https://doi.org/10.30646/sinus.v19i1.526

Barburiceanu, S., Meza, S., Orza, B., & Malutan, R. (2021). Convolutional Neural Networks for Texture Feature Extraction . Applications to Leaf Disease Classification in Precision Agriculture. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2021.3131002

De Silva, M., & Brown, D. (2023). Multispectral Plant Disease Detection with Vision Transformer–Convolutional Neural Network Hybrid Approaches. *Sensors*, *23*(20). https://doi.org/10.3390/s23208531

Dubey, R. K., & Choubey, D. K. (2024). RETRACTED ARTICLE: Feature selection with Optimized XGBoost model-based paddy plant leaf disease classification. *Multimedia Tools and Applications*, *83*, 80281.

https://api.semanticscholar.org/CorpusID:267990380

Hutauruk, A. R. (2025). *Smart Rice Disease Detection Based on Leaf Analysis Using the YOLO Algorithm with an Interactive User Interface*. *08*(02), 188–190.

Jordy, R., & Ariatmanto, D. (2025). *Perbandingan Metode Ekstraksi Fitur LBP , GLCM , dan Canny dalam Klasifikasi Penyakit Daun Padi dengan KNN*. *14*(02), 44–51. https://doi.org/10.52771/bangkitindonesia.v14i2.452

Khalil, W., Irsan, M., & Fathoni, M. F. (2024). *Designing an Application for Detecting Diseases of Rice Plants Using OOAD Method*. *8*(2), 974–982.

Khoiruddin, M., & Tena, S. (2024). *Fruit and Vegetable Classification using Convolutional Neural Network with MobileNetV2*. *2*(2), 203–210. https://doi.org/10.61098/jarcis.v2i2.197

Kulkarni, P., & Shastri, S. (n.d.). *Registered under MSME Government of India Rice Leaf Diseases Detection Using Machine Learning*. *10*, 17–22.

Kusanti, J., Penyakit, K., Padi, D., & Haris, A. (2018). Klasifikasi Penyakit Daun Padi Berdasarkan Hasil Ekstraksi Fitur GLCM Interval 4 Sudut. *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, *03*(01), 1–6.

Miftahushudur, T., Sahin, H. M., & Grieve, B. (2025). *A Survey of Methods for Addressing Imbalance Data Problems in Agriculture Applications*. 1–32.

Milano, A. C., Yasid, A., & Wahyuningrum, R. T. (2024). *KLASIFIKASI PENYAKIT DAUN PADI MENGGUNAKAN MODEL DEEP LEARNING EFFICIENTNET-B6*. *12*(1).

Nata, H., Pirnando, N., & Petrus, J. (2025). *Klasifikasi Penyakit Daun Padi Menggunakan Convolutional Neural Network dengan Arsitektur AlexNet*. 207–214.

Priyangka, A. A. J. V., & Kumara, I. M. S. (2021). Classification Of Rice Plant Diseases Using the Convolutional Neural Network Method. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *12*(2), 123. https://doi.org/10.24843/lkjiti.2021.v12.i02.p06

Ramli, R., & Riadi, A. A. (2025). *Classification of Rice Leaf Diseases Using Support Vector Machine with HSV and GLCM-Based Feature Extraction*. *9*(5), 2329–2337.

Seelwal, P., Dhiman, P., Gulzar, Y., Kaur, A., Wadhwa, S., & Onn, C. W. (2024). *A systematic review of deep learning applications for rice disease diagnosis : current trends and future directions*. *September*. https://doi.org/10.3389/fcomp.2024.1452961

Sharma, R., Singh, A., Kavita, Jhanjhi, N. Z., Masud, M., Jaha, E. S., & Verma, S. (2021). Plant Disease Diagnosis and Image Classification Using Deep Learning. *Computers, Materials and Continua*, *71*(2), 2125–2140. https://doi.org/https://doi.org/10.32604/cmc.2022.020017

Shrivastava, V. K., & Pradhan, M. K. (2021). Rice plant disease classification using color features: a machine learning paradigm. *Journal of Plant Pathology*, *103*(1), 17–26. https://doi.org/10.1007/s42161-020-00683-3

Sovia, N. A., Wayan, N., Wardhani, S., & Sumarminingsih, E. (2025). *Enhancing Image Classification of Cabbage Plant Diseases Using a Hybrid Model Convolutional Neural Network and XGBoost*. *10*(1), 278–289.

Talreja, R., Jawrani, V., Watwani, B., Sengupta, S., Rohera, P., & Raghuwanshi, K. (2022). *AgriCare: An Android Application for Detection of Paddy Diseases*. 1–6. https://doi.org/10.1109/INCET54531.2022.9825038

Tiwari, R., Patel, J., Khan, N. R., & Dadhich, A. (2025). *automated identification of rice diseases*. 1–16. https://doi.org/10.1371/journal.pone.0307461

Wibisono, N. B., & Saiful, S. (2025). *Crop Yield Prediction Using Random Forest Algorithm and Xgboost Machine Learning Model*. *IX*(2454), 1983–1994. https://doi.org/10.47772/IJRISS

Widyawati, W., AR, N. H., Syafrial, S., & Sujarwo, S. (2025). Crafting the future of rice in Indonesia: sustainable supply through systems thinking. *Cogent Social Sciences*, *11*(1), 2488113. https://doi.org/10.1080/23311886.2025.2488113