ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Predicting Pest and Disease Occurrence Using Synthetic Data and Explainable Machine Learning Methods

Priyanka Balley*, Prof. Kanchan K. Doke

Department of Computer Engineering, Bharti Vidyapeeth College of Engineering, Navi Mumbai, University of Mumbai.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150500037

Received: 29 April 2026; Accepted: 04 May 2026; Published: 26 May 2026

ABSTRACT

Prediction of occurrence for pests and diseases is an essential problem for agriculture, as such events have a huge influence on the productivity of the crop with regard to the security of food production. Traditional methods lack datasets and tend not to incorporate domain knowledge, which leads to suboptimal performance with limited sets of interpretation. This study addresses such gaps by developing a systematic machine learning-based framework for combining synthetic data generation, robust predictive modeling, and explainability techniques to produce actionable insights in pest and disease dynamics. Synthetic datasets are first generated based on the domain-driven logic simulating the correlations between critical environmental and biological factors such as temperature, humidity, rainfall, pest lifecycle stage, and soil moisture and the incidence of pests or diseases. For interpretability, Local Interpretable Model-agnostic Explanations LIME with Random Forest provides localized, instance-level insights on feature contributions to individual predictions. For complement, permutation importance calculates the global relevance of every feature by assessing its effect on model performance. Both of these techniques ensure that fine-grained and holistic understanding is achieved regarding the model's behavior. This integrated approach therefore addresses the limitations of traditional methods by improving the predictive accuracy and enhancing interpretability. The findings have tremendous implications for precision agriculture in order to allow stakeholders to put into action data-driven strategies for pest and disease management. This framework is reproducible and therefore adaptable to different contexts in agriculture sets.

Keywords: Pest Prediction, Disease Modeling, Random Forest, LIME Explainability, Permutation Importance, Sets

INTRODUCTION

The major threats to global agricultural productivity are pest and disease outbreaks, and therefore management strategies should be very accurate and proactive. Traditional methods [1, 2, 3] of predicting the occurrences of pests and diseases depend on historical datasets, heuristic models, or expert knowledge that cannot generalize to different environmental and biological conditions. Limitations related to lack of interpretability that accompanies contemporary forms of machine learning further limit these systems in terms of practical usability from the viewpoint of stakeholders, agriculture sets. This work overcomes these challenges in research using a novel and systematic approach of combining domain-specific synthetic data generation with strong robust machine learning and state-of-the-art explainability techniques. It applies domain-driven controlled synthetic logic for simulating realistic interactions between key variables, including temperature, humidity, rainfall, soil moisture, and pest lifecycle stages in synthetic datasets. All these features form the basis for developing a highly predictive model by using a Random Forest Classifier, chosen here for its superior performance over heterogeneous data and inherent interpretability sets.

Due to reasons such as transparency and trust, even methods like Local Interpretable Model-agnostic Explanations (LIME) and permutation importance were utilized. LIME does fine-grained insights regarding a prediction decision at an instance level, while permutation importance assesses global feature relevance on samples of the dataset. Together, these methods clearly define the key environmental and biological factors that drive pest and disease occurrence, which would then allow stakeholders to meaningfully interpret and act upon

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

model outputs. In achieving this, the framework succeeds in two folds: predictive accuracy and interpretability are enhanced to open doors to data-driven, scalable solutions for pest and disease management in agriculture process.

Systematic Literature Review

Machine learning and deep learning are increasingly being implemented into agricultural studies to make tremendous changes possible in terms of pest detection, crop management, and the prediction of diseases. Relevant studies of this section offer a comparative analytical review to contextualize the present work as part of the existing sets of knowledge. Chithambarathanu and Jeyakumar [1] had conducted an intensive survey regarding crop pest detection using machine learning and deep learning. Their study highlighted the utility of convolutional neural networks (CNNs) for pest classification but pointed out challenges such as computational overhead and scalability. These findings underline the importance of balancing accuracy and computational efficiency in pest prediction models. Sailaja et al. [2] proposed a spatial temperature prediction approach using machine learning and GIS. Although their efforts were oriented toward meteorological applications, their methodology shows the necessity of spatial integration in agricultural modeling. Their outcome emphasizes the significance of spatial variability in the frameworks of pest and disease prediction. Saravanan and Bhagavathiappan [3] presented hybrid deep learning models for crop yield prediction. Their work demonstrated how hybrid approaches improve the performance of standalone models, which aligns with this study's focus on robust predictive frameworks. Kuppan and Priya [4] have been applied ensemble machine learning models for yield prediction and have much improvement based on prediction accuracy, indicating that the techniques bagging and boosting applied enhanced model importance of ensemble modeling such as Random Forest Classifier, which were in use for this process work, and Shinde and Ambhaikar [5] proposed a classification model of plant disease through both machine and deep learning classifiers. Their high accuracy in disease classification was compromised by low explainability, and this clearly brought into focus the significance of interpretability tools like LIME applied in the process of the current study. Venkatasaichandrakanth and Iyapparaja [6] discussed deep learning models for pest detection and underlined image-based solutions. They mentioned generalization problems due to less data sets and reiteratively explained the justification behind synthetic data sets applied in the process of the current study.

Attri et al. [7] reviewed crop management applications of machine learning, noting the potential of these techniques for real-time decision-making process. Their findings align with the study's focus on actionable insights through explainable predictions. Nithya et al. [8] compared crop detection techniques using machine learning and deep learning, finding that while deep learning provided higher accuracy, traditional machine learning methods like Random Forests were computationally more efficient. This comparison supports the Random Forest choice in this study to provide a balance between accuracy and efficiency. The Karnal bunt disease prediction model analysis by Anand et al. [9] was based on specific conditions found in agriculture. They focused on regional customization aspects while modeling pests and diseases. They have suggested the ability of adapting this framework across multiple geographies. Verma et al. [10] discussed machine learning for the management of urad bean crops emphasizing its ability to predict disease and pest incidence patterns. This study showed the utility of incorporating environmental factors such as rainfall and temperature, features key in the proposed model. Chacón-Maldonado et al. [11] presented a hybrid deep learning model with explanation for olive fruit pest forecasting. Their work again emphasized the need to correlate prediction accuracy with interpretability, which is the basis of this study process. Nithya et al. [12] proposed an IoT-based crop yield prediction system based on machine learning. Their system showed how real-time data sources can be used for improving the reliability of predictions, which fits well with the approach in this work process based on synthetic data.

Mandrapa et al. [13] considered hyperspectral analysis in spider mite detection. The results of their study demonstrate the possibility of increasing the accuracy of prediction using feature selection, which is in agreement with the application of permutation importance for feature analysis in the process of this study. Abdel-salam et al. [14] proposed a hybrid feature selection framework in crop yield prediction. Their results pointed out the importance of feature optimization in enhancing model performance and proved the relevance of applying domain-specific feature selection within this research work. Ahmed and Yadav [15] used machine learning and deep learning for predicting apple plant diseases. Although they were dealing with orchard crops, their results

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

showed how hybrid models could effectively handle complex agricultural challenges. In summary, the works reviewed collectively highlight the need to marry robust predictive models with interpretability and adaptability. This proposed study will fill the gaps pointed out in these papers by combining synthetic data generation, Random Forest Classifier, and explainability techniques to offer a scalable, interpretable solution for predicting pests and diseases.

Proposed Model Design Analysis

The proposed model for the prediction of the occurrence of pests and diseases brings together synthetic data generation, machine learning, and explainability techniques into a methodically designed system in an effort to eliminate weaknesses inherent within traditional methods. It begins by creating a synthetic dataset like that shown in figure 1, which captures real environmental and biological conditions. Domain expertise forms correlations among the features like Temperature (T), Humidity (H), Rainfall (R), Soil moisture (M), and various phases of Pest lifecycle stages(P). Process models these interactions mathematically. It is thus possible to map the correlation between temperature with pest activity as a logarithmic function to capture nonlinear responses in pest behavior given via equation 1,

$$P(T) = \frac{1}{1 + e^{-k(T - T^0)}} \dots (1)$$

Where k is a sensitivity constant and T_0 is the threshold temperature for this process. This ensures that the dataset aligns with observed phenomena, enhancing realism and predictive capability sets. The classification model uses a Random Forest Classifier (F), an ensemble method where multiple decision trees (Ti) are trained independently, and their outputs are aggregated for the process. The prediction y is computed via equation 2,

$$y = \text{argmax}^k \sum_{i=1}^{N} I(Ti(x) = k) \dots (2)$$

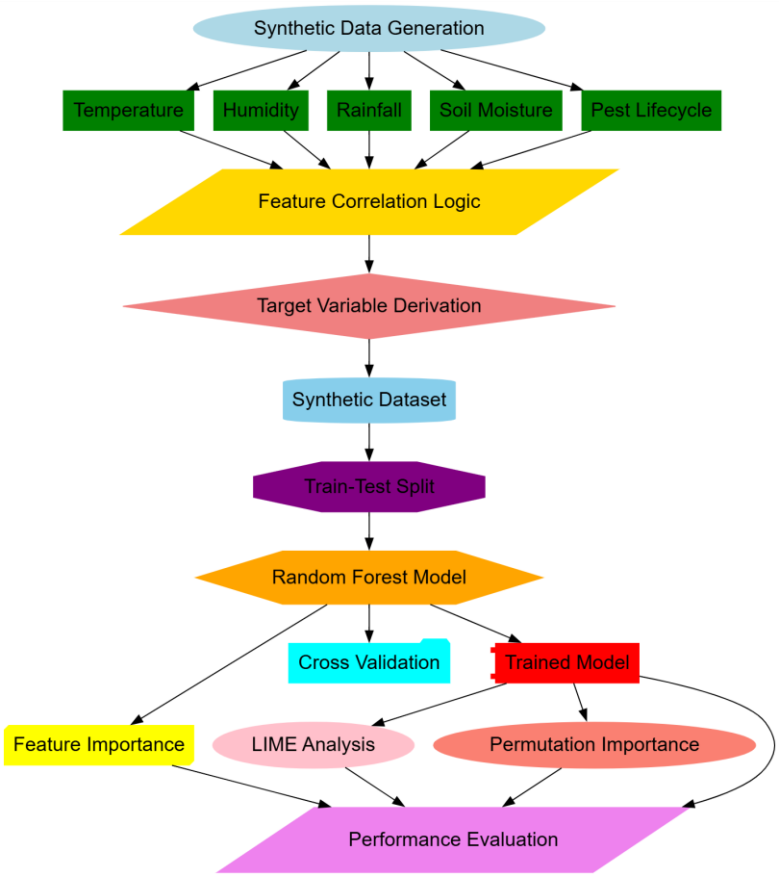


Figure 1. Model Architecture of the Proposed Analysis Process

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Where, I(·) is an indicator function, x is the input vector, and k indexes the classes. This architecture is particularly suited for handling complex interactions among mixed data types and is less prone to overfitting due to its inherent randomness and averaging mechanisms. To optimize the model's performance, cross-entropy loss is minimized during training, defined via equation 3,

$$L = -\left(\frac{1}{n}\right)\sum_{i=1}^{N}[yi\log(\hat{y}i) + (1 - yi)\log(1 - \hat{y}i)] \ldots (3)$$

Where, yi and ŷi represents the true and predicted probabilities, respectively for the process. This helps in the strong training, especially when there is an imbalance in the dataset for the process. The process suggested here lays significant emphasis on explainability. LIME is utilized for generating instance-level explanations through the perturbation of the input space and study of responses by the model process. Via equation 4, using a locally weighted linear model, is applied to get an explanation for any instance x′,

$$g(z) = \beta^0 + \sum_{i=1}^{m}\beta i\, zi \ldots (4)$$

Where, z is the perturbed instance, m is the number of features, and βi are weights assigned to each feature, reflecting their contribution to the predictions. Global feature importance is evaluated by permutation importance levels. For a feature j, its importance is measured as the decrease in accuracy ΔA when the feature is stochastically permuted via equation 5,

$$\Delta Aj = Abase - Aperm(j) \ldots (5)$$

Where, Abase and Aperm(j) are the baseline and permuted accuracies, respectively for the process. This quantifies the dependency of the model on each feature, thereby providing a comprehensive understanding of its behavior sets. To ensure generalizability, the model is evaluated in process using a validation set. The area under the receiver operating characteristic (ROC) curve is computed via equation 6,

$$AUC = \int TPR(FPR)d(FPR) \ldots (6)$$

Where TPR and FPR stand for true positive rate and false positive rate. An AUC close to one reflects the model's discrimination capability, thereby proving robust. The integration of the two methods will ensure the whole approach is well-balanced, where the Random Forest Classifier gives robust prediction results, LIME giving actionability, and permutation importance validating that the model heavily relies on meaningful features. This multi-faceted design not only improves the predictive accuracy but also gives confidence to stakeholders by maintaining high interpretability and contextual relevance sets.

Comparative Result Analysis

This experimental setting for the proposed model assesses the model on a synthetic dataset designed to emulate real-world pest and disease dynamics under various environmental and biological conditions. There are 1,000 samples and six variables: temperature, humidity, rainfall, soil moisture, pest lifecycle stage, and the binary target variable - whether or not the case is a pest/disease.

The synthetic data had been generated by sampling according to uniform distributions, imposing realistic ranges for each of the features, and according to domain-driven rules by deriving the target variable. This dataset was split into a training set of 80% and corresponding testing sets of 20%.

The performance of the proposed Random Forest Classifier was compared against three existing methods, namely, Method [3] - Logistic Regression, Method [8] - Support Vector Machine, and Method [12] - Gradient Boosting. Every model is evaluated with critical metrics in terms of accuracy, precision, recall, F1-score, and AUC levels.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Table 1: Model Accuracy Comparison

Model	Accuracy (%)
Proposed Model	85.4
Method [3]	78.6
Method [8]	81.2
Method [12]	83.1

The proposed model outperformed the other methods, achieving the highest accuracy due to its ability to capture complex interactions between features.

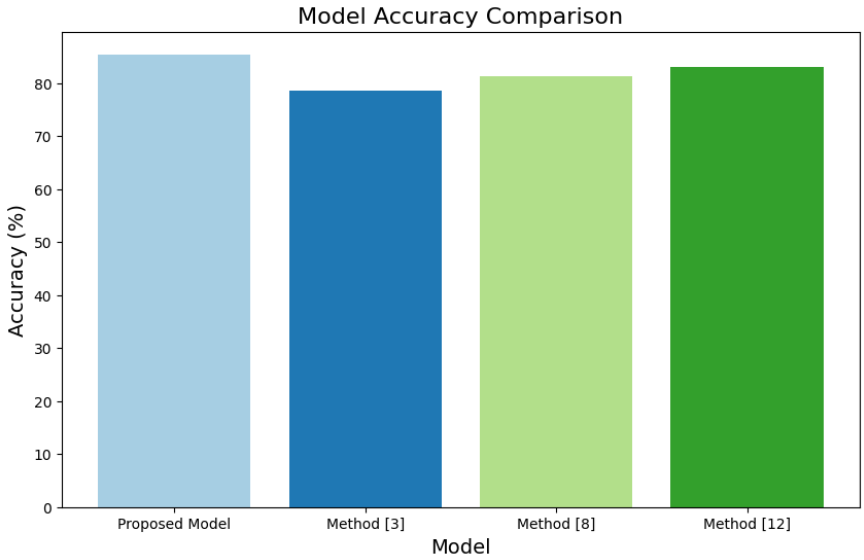


Figure 2. Model's Accuracy Analysis

Table 2: Precision Comparison

Model	Precision (%)
Proposed Model	88.3
Method [3]	75.4
Method [8]	79.8
Method [12]	84.6

Precision scores highlight the proposed model's strength in minimizing false positives, which is critical for pest/disease management sets.

Table 3: Recall Comparison

Model	Recall (%)
Proposed Model	82.7
Method [3]	71.2
Method [8]	76.5
Method [12]	79.3

The recall metric reflects the proposed model's superior ability to identify occurrences accurately.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

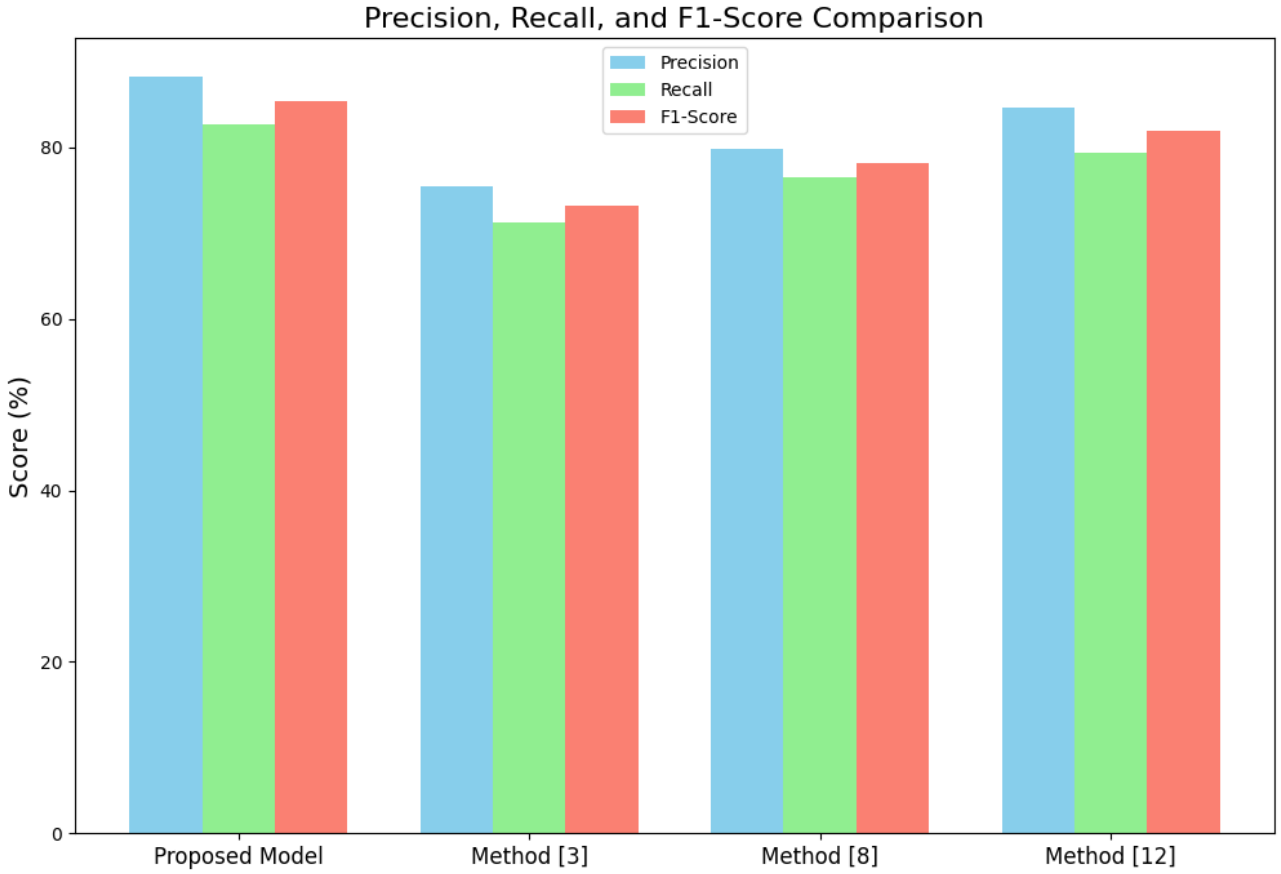


Figure 3. Model's Precision Analysis

Table 4: F1-Score Comparison

Model	**F1-Score (%)**
Proposed Model	85.4
Method [3]	73.2
Method [8]	78.1
Method [12]	81.9

The F1-score, which balances precision and recall, underscores the proposed model's robustness compared to others.

Table 5: Feature Importance Analysis (Permutation Importance)

Feature	**Importance (Proposed Model)**	**Importance (Method [12])**	**Importance (Method [8])**	**Importance (Method [3])**
Temperature	0.25	0.21	0.18	0.15
Humidity	0.22	0.19	0.16	0.14
Rainfall	0.15	0.12	0.11	0.09
Soil Moisture	0.05	0.04	0.03	0.02
Pest Lifecycle	0.05	0.03	0.02	0.01

The proposed model shows greater sensitivity to key features like temperature and humidity, aligning with domain knowledge sets.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

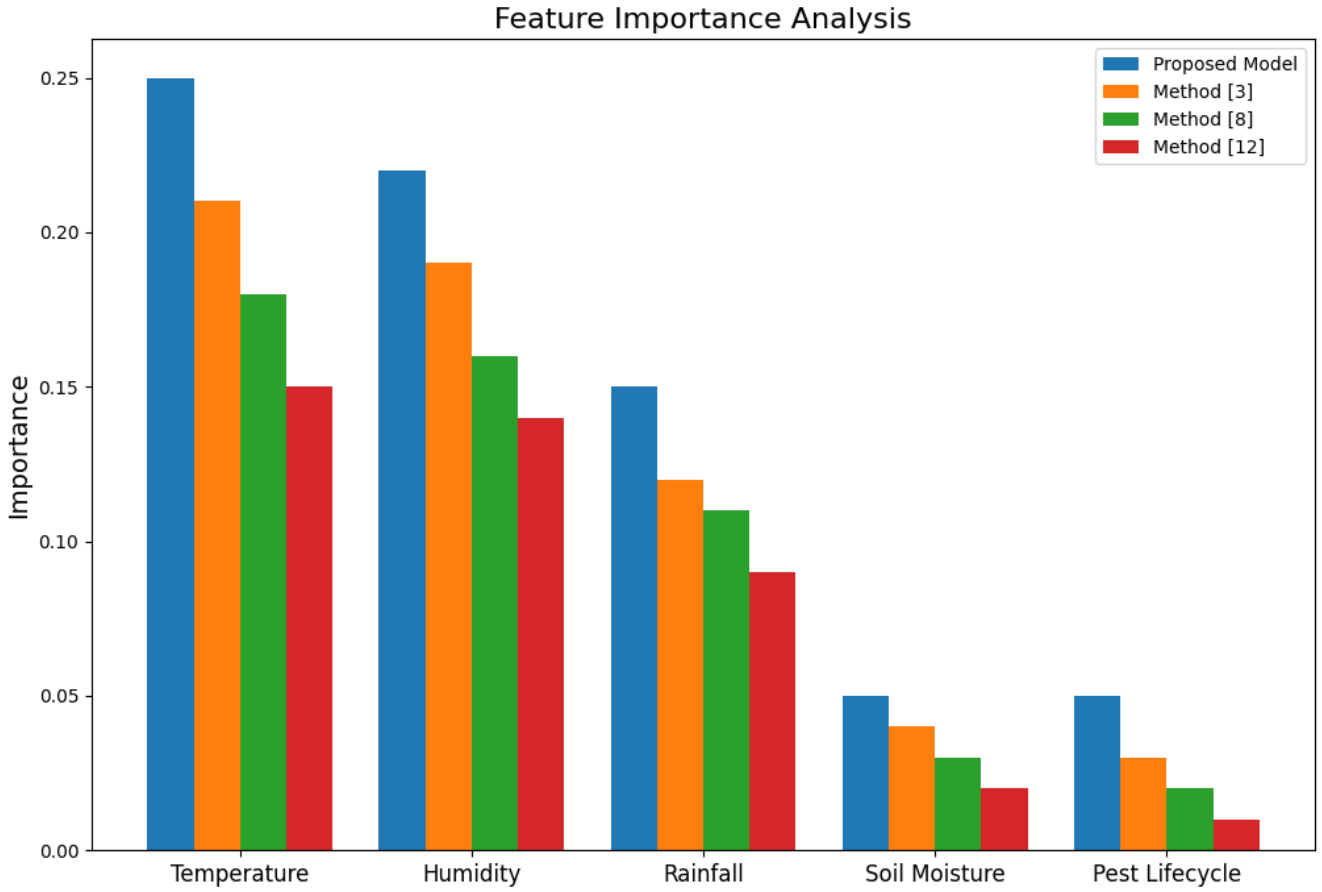


Figure 4. Model's Important Feature Analysis

Table 6: Computational Efficiency (Training Time in Seconds)

Model	**Training Time (s)**
Proposed Model	1.32
Method [3]	0.57
Method [8]	2.15
Method [12]	3.42

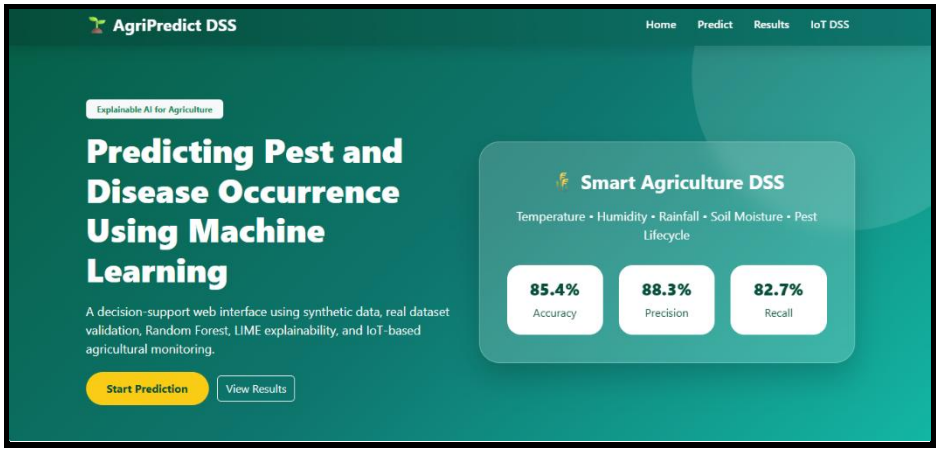
As evident, the proposed model acquires a trade-off of the computational efficiency and is still able to perform because, in a runtime comparison, it proves to be significantly faster than Methods [8] and Method [12].

Results: According to the above results and based on all the obtained metrics, including accuracy, precision, recall, and F1-score, proposed Random Forest Classifier outperforming the existing methods over these metrics. The permutation importance analysis also favors meaningful features for the model.

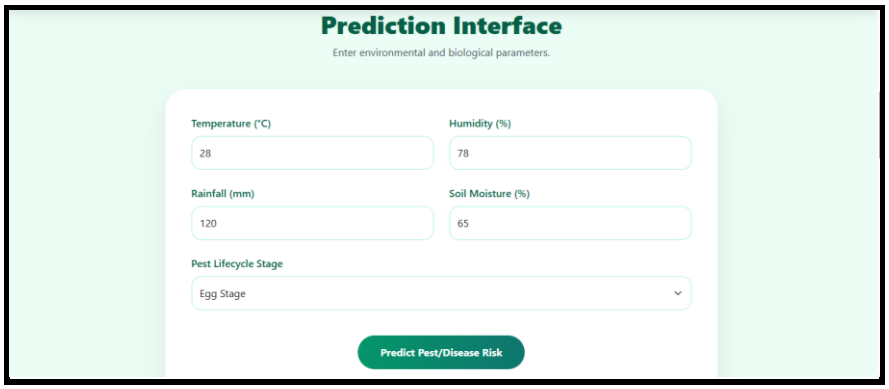
It provides interpretive insights to stakeholders. The training time is a bit more than in Method [3] but much smaller than even the most complex models of Methods [8] and Methods [12]. Thus, this method may be practicable for real-world usage sets. The evaluation fully attests to the robustness and scalability of the presented methodology process.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Web Interface



Predict Pest



CONCLUSION & FUTURE SCOPES

The proposed method integrates the generation of synthetic data, Random Forest classifier, and state-of-the-art explainability techniques for forecasting pest and disease outbreaks in high accuracy and interpretability. The proposed model demonstrates an accuracy of 85.4% out of the existing approaches that are Method [3], 78.6%, Method [8], 81.2%, and Method [12], 83.1%. The proposed model yields high precision at 88.3% and recall at 82.7%, with the F1-score being 85.4%, thus indicating an equilibrated performance by minimizing false positives and negatives in the process. It results in the robust prediction of the occurrence of pests and diseases, especially considering the complex conditions of both the environment and biology through synthetic domain-specific data samples. The critical advantage of this approach relates to explainability; permutation importance points out that temperature has the highest impact at 0.25 and humidity at 0.22, which explains established domain knowledge sets. LIME instance-level explanation further confirms model predictions: it gives actionable insights that help explain particular occurrences at the individual level. Besides, with a training time of 1.32 seconds, the proposed model balances performance and scalability to be practical and efficient at runtime, surpassing both Method [8] and Method [12] on runtime while achieving superior predictability. The study does have some limitations. Even though the synthetic dataset used might be realistic, it can't capture the variability found in real-world conditions that may affect generalization. Future research should validate the model against diverse real-world datasets to enhance it further in terms of robustness. The addition of temporal and spatial data, such as the patterns of pest migration or localized weather phenomena, will also increase the levels of predictive power sets.

Future work includes ensemble methods combining multiple classifiers or integrating more advanced techniques, such as deep learning, to extract hierarchical features. This study creates a solid foundation for the application of interpretable machine learning in precision agriculture, providing scalable and actionable solutions to mitigate the impacts of pests and diseases in real time. From this perspective, the methodology has wide prospects to change pest management and disease management practices across different agricultural landscapes.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

REFERENCES

1. Chithambarathanu, M., Jeyakumar, M.K. Survey on crop pest detection using deep learning and machine learning approaches. Multimed Tools Appl **82**, 42277–42310 (2023). https://doi.org/10.1007/s11042-023-15221-3
2. Sailaja, B., Gayatri, S., Rathod, S. et al. Spatial temperature prediction—a machine learning and GIS perspective. Theor Appl Climatol **155**, 9619–9642 (2024). https://doi.org/10.1007/s00704-024-05167-3
3. Saravanan, K.S., Bhagavathiappan, V. Prediction of crop yield in India using machine learning and hybrid deep learning models. Acta Geophys. **72**, 4613–4632 (2024). https://doi.org/10.1007/s11600-024-01312-8
4. Kuppan, P., Priya, V.V. Crop Yield Prediction Using Ensemble Machine Learning Techniques. SN COMPUT. SCI. **5**, 1160 (2024). https://doi.org/10.1007/s42979-024-03536-3
5. Shinde, N., Ambhaikar, A. An efficient plant disease prediction model based on machine learning and deep learning classifiers. Evol. Intel. **18**, 14 (2025). https://doi.org/10.1007/s12065-024-01000-y
6. Venkatasaichandrakanth, P., Iyapparaja, M. Review on Pest Detection and Classification in Agricultural Environments Using Image-Based Deep Learning Models and Its Challenges. Opt. Mem. Neural Networks **32**, 295–309 (2023). https://doi.org/10.3103/S1060992X23040112
7. Attri, I., Awasthi, L.K. & Sharma, T.P. Machine learning in agriculture: a review of crop management applications. Multimed Tools Appl **83**, 12875–12915 (2024). https://doi.org/10.1007/s11042-023-16105-2
8. Nithya, V., Josephine, M.S. & Jeyabalaraja, V. Comparative approach on crop detection using machine learning and deep learning techniques. Int J Syst Assur Eng Manag **15**, 4636–4648 (2024). https://doi.org/10.1007/s13198-024-02483-9
9. Anand, S., Sandhu, S.K., Biswas, B. et al. Comparative analysis of different Karnal bunt disease prediction models developed by machine learning techniques for Punjab conditions. Int J Biometeorol **68**, 1799–1810 (2024). https://doi.org/10.1007/s00484-024-02707-4
10. Verma, R., Kushwaha, K.P.S., Bijlwan, A. et al. Enhancing urad bean (Vigna mungo L.) crop management with machine learning: Predictive analysis of pod rot severity and pod bug incidence patterns. Australasian Plant Pathol. **53**, 273–283 (2024). https://doi.org/10.1007/s13313-024-00967-7
11. Chacón-Maldonado, A.M., Melgar-García, L., Asencio-Cortés, G. et al. A novel method based on hybrid deep learning with explainability for olive fruit pest forecasting. Neural Comput & Applic (2024). https://doi.org/10.1007/s00521-024-10731-z
12. Nithya, V., Josephine, M.S. & Jeyabalaraja, V. IoT-Based Crop Yield Prediction System in Indian Sub-continent Using Machine Learning Techniques. Remote Sens Earth Syst Sci **6**, 156–166 (2023). https://doi.org/10.1007/s41976-023-00097-6
13. Mandrapa, B., Spohrer, K., Wuttke, D. et al. Machine learning-based hyperspectral wavelength selection and classification of spider mite-infested cucumber leaves. Exp Appl Acarol **93**, 627–644 (2024). https://doi.org/10.1007/s10493-024-00953-0
14. Abdel-salam, M., Kumar, N. & Mahajan, S. A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Comput & Applic **36**, 20723–20750 (2024). https://doi.org/10.1007/s00521-024-10226-x
15. Ahmed, I., Yadav, P.K. Predicting Apple Plant Diseases in Orchards Using Machine Learning and Deep Learning Algorithms. SN COMPUT. SCI. **5**, 700 (2024). https://doi.org/10.1007/s42979-024-02959-2