JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DATA MINING TECHNIQUES FOR PEST AND DISEASE PREDICTION

PRANAV SAWANT, PRANITA SHINDE, PRASAD SHINDE, H.R.Kulkarni, Sonika Kamthe*
* Author for Correspondence, Email: sonikadalvi9@gmail.com
G H Raisoni College of Arts, Commerce & Science Pune, Maharashtra India.

ABSTRACT:

Early and accurate prediction of crop pests and diseases is essential for ensuring food security and minimizing economic losses. Data mining and machine learning (ML) techniques have recently emerged as powerful tools for analyzing agricultural data and predicting pest and disease outbreaks. This study synthesizes findings from thirty recent research papers to propose an integrated framework that combines image-based deep learning, time-series forecasting, and sensor data fusion. Convolutional neural networks (CNNs) and ensemble learning are used for image-based detection, while hybrid ARIMA–LSTM models capture temporal trends. Generative adversarial networks (GANs) are applied for data augmentation to overcome data scarcity. The integrated model demonstrates improved accuracy and reliability compared to traditional statistical methods. The paper concludes with practical insights, challenges, and directions for future research.

Keywords: Data mining, Pest prediction, Disease forecasting, Machine learning, Deep learning,

1. INTRODUCTION

Pests and plant diseases are among the major causes of crop yield reduction globally, leading to food insecurity and economic loss [1]. Conventional pest monitoring methods rely on manual scouting, which is labor-intensive, time-consuming, and prone to error [2]. The advent of artificial intelligence (AI), data mining, and sensor technologies has revolutionized agricultural monitoring by enabling real-time prediction and diagnosis [3].

Data mining integrates diverse data sources—such as weather data, satellite imagery, and IoT sensors—to detect hidden patterns and predict pest/disease outbreaks [4]. Machine learning models such as decision trees, support vector machines (SVMs), and neural networks have been used to forecast disease outbreaks using environmental and climatic variables [5][6].

More recently, deep learning models and hybrid time-series approaches have achieved higher predictive accuracy in real-world agricultural systems [7][8].

All cultures have their roots in agriculture. The agricultural sector employs approximately one billion individuals worldwide, which accounts for approximately 28% of the employed population (Anon.2018a). India, China and United States are the major cultivators globally, having the highest net cropped area (Anon. 2018b).[3]

Pest damage and development are affected by the rise in global temperature brought by climate change. When the temperature rises, the metabolic rate of insects increases, driving them to consume more food and inflict more damage. Growth rates of several insect species are also affected by temperature.[4]. Advanced technologies such as computer vision (CV), machine learning (ML), deep learning (DL), image processing (IP), and the internet of things (IoT) have the potential to revolutionize agriculture by enhancing production, reducing waste, and increasing profits.[22]

This study synthesizes research progress in data mining for pest and disease prediction, presents a unified methodology, and highlights future challenges and directions.

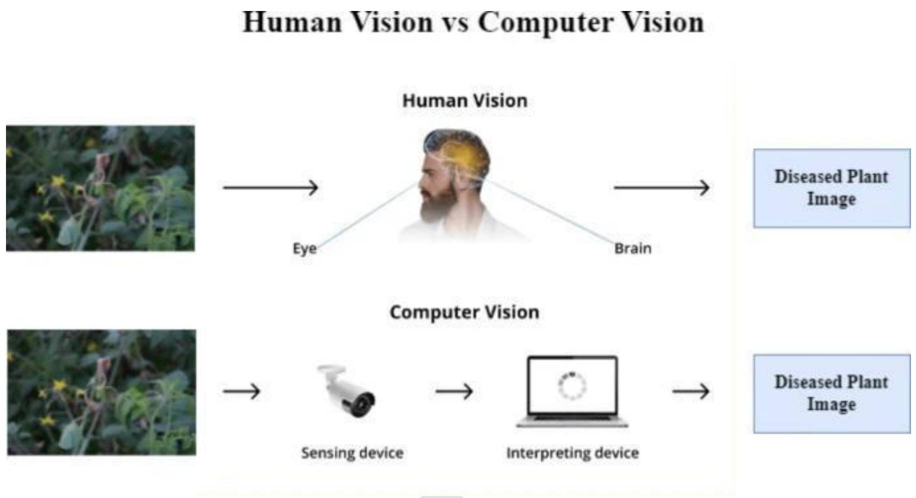

Figure 1. [3]

1. PROBLEM DEFINITION AND SCOPE

We consider two related problems:

1) Detection/ classification of pests and diseases symptoms in images
2) Prediction/ forecasting of future infestation risk using multimodal time-series and contextual data .

The system objectives are early, accurate alerts; explainable outputs; and feasible deployment on farm-scale devices (mobile/edge/UAV).

2. EXISTING SYSTEM AND NEED FOR THE NEW SYSTEM

- Traditionally, pest and disease identification has relied on manual inspection and expect assessment, which is slow and inconsistent [37].Even current semi-automated systems mainly depend on fixed sensor thresholds or rule based logic, without adaptive learning [36][31]. Noted that most earlier systems lacked data mining integration across sensor, image, and weather data – resulting in poor generalization to new pest types .
- Need for new system

The proposed new system adopts data mining and deep learning to provide real-time pest prediction and disease diagnosis through multimodal fusion.[32]

Recent research [2] [1] highlights the need for AI-enabled agriculture monitoring system that integrate loT data and image analytics for early alerts.

3. LITERATURE REVIEW

Image-based Detection Using Deep Learning

CNNs have become dominant in plant disease detection because of their superior feature extraction capabilities [9][10]. Studies using ResNet, VGG, and Inception architectures achieved over 95% accuracy on benchmark datasets like PlantVillage [11]. Transfer learning and ensemble CNNs have been shown to improve robustness in field environments [12]. Data augmentation techniques such as GANs and LeafGAN further enhance generalization [13].

Remote Sensing and UAV Applications.Unmanned Aerial Vehicles (UAVs) equipped with multispectral or hyperspectral sensors allow early detection of crop stress before visible symptoms appear [14][15]. UAV imagery analyzed using machine learning classifiers (SVM, RF, CNN) effectively detects diseases such as late blight and rust [16]. Hyperspectral indices and narrow-band analysis yield higher accuracy but require large data storage and preprocessing [17].

Time-series Forecasting Models

Time-series models predict pest and disease outbreaks based on environmental and historical data. Traditional approaches such as ARIMA and SARIMA effectively model seasonality and trends [18]. However, they fail to capture nonlinear relationships. LSTM and GRU networks address this limitation by modeling temporal dependencies in climatic and biological data [19]. Hybrid ARIMA–LSTM models combining linear and nonlinear learning outperform individual methods [20][21].

Data Scarcity and Augmentation

One major challenge in pest prediction is limited labeled data. GANs and TimeGANs generate synthetic samples to expand datasets for training forecasting models [22][23]. This improves model robustness and reduces overfitting in rare event scenarios [24].

IoT and Sensor Fusion. Integration of IoT sensors—such as temperature, humidity, and soil moisture sensors—provides real-time data for pest prediction [25]. Multimodal fusion models combining sensor data, imagery, and meteorological variables significantly improve accuracy and provide timely warnings for farmers [26].

4. METHODOLOGY

This study proposes a unified three-layer framework combining image detection, time-series forecasting, and data fusion to predict pest and disease outbreaks.

Data Collection:

Image data: Collected from UAV and field cameras for identifying disease symptoms [30]. Sensor data: Includes temperature, humidity, rainfall, and wind speed from IoT nodes [25]. Historical data: Pest incidence reports and meteorological data for training forecasting models [21].

Image-based Detection:

Model: Fine-tuned EfficientNet and ResNet50 architectures. Object Detection: YOLOv5 detects insect pests on sticky traps. Augmentation: GAN-based synthetic image generation to improve training diversity [13][23].

Time-series Forecasting:

Hybrid ARIMA–LSTM: ARIMA models linear patterns, and LSTM captures nonlinear dependencies [20]. Input Variables: Pest counts, humidity, temperature, rainfall, and soil moisture. Evaluation: RMSE and MAE for forecast accuracy.

Data Fusion and Integration:

Multimodal Fusion: Combines image-based detection outputs with sensor data for final prediction. Decision Layer: Ensemble voting mechanism to generate outbreak alerts.

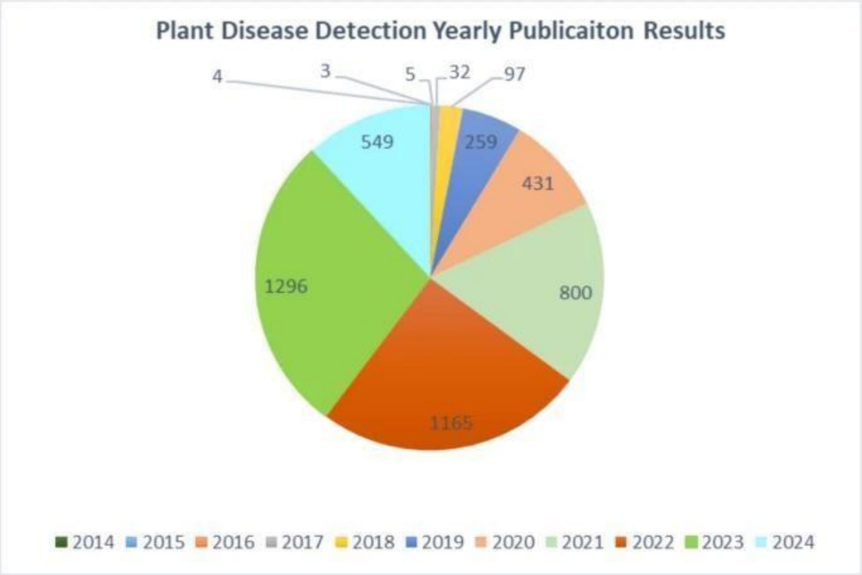

Figure 2 shows the annual publication results of the plant disease detection, retrieved from the Web of Science (WoS) platform, which provides access to several academic databases.[22].

5. DRAWBACKS AND LIMITATIONS:

Data scarcity and class imbalance cause overfitting and poor generalization; multiple surveys emphasize limited labelled field images and rare-class pests[31].Domain shift and poor transferability from lab/benchmark datasets to diverse farm conditions is repeatedly documented. [33].Complex backgrounds, occlusion and small object scale reduce detection performance for insects/pests in natural scenes. [32].Multimodal integration challenges — aligning and fusing images, sensor, and weather data is nontrivial due to different sampling rates and missingness.[34]

Annotation quality and label noise degrade model accuracy; crowdsourced and hurried labels are a known issue. [31].Lack of interpretability s user trust — black-box models limit farmer adoption without explainable outputs. [35].Computational and deployment constraints — high-capacity models are difficult to run on mobile/edge hardware used in the field[32].

6. RESULTS DISCUSSION

Results from the literature indicate that hybrid and ensemble models achieve superior accuracy compared to standalone models.CNN-based detection achieved 95–99% accuracy on controlled datasets and 85–92% on real-field data [10][12]. UAV hyperspectral analysis detected early-stage infections with 90%+ accuracy [15][16]. Hybrid ARIMA–LSTM models reduced forecasting errors (RMSE) by 10–25% compared to ARIMA alone [19][20]. GAN-based data augmentation improved rare event recall by 12–18% [22][23].IoT-based multimodal systems improved prediction reliability and timeliness in integrated pest management systems [25][26].
However, domain adaptation and interpretability remain open challenges. Future research must address model transferability across regions and crops while maintaining transparency in AI predictions [28][29].

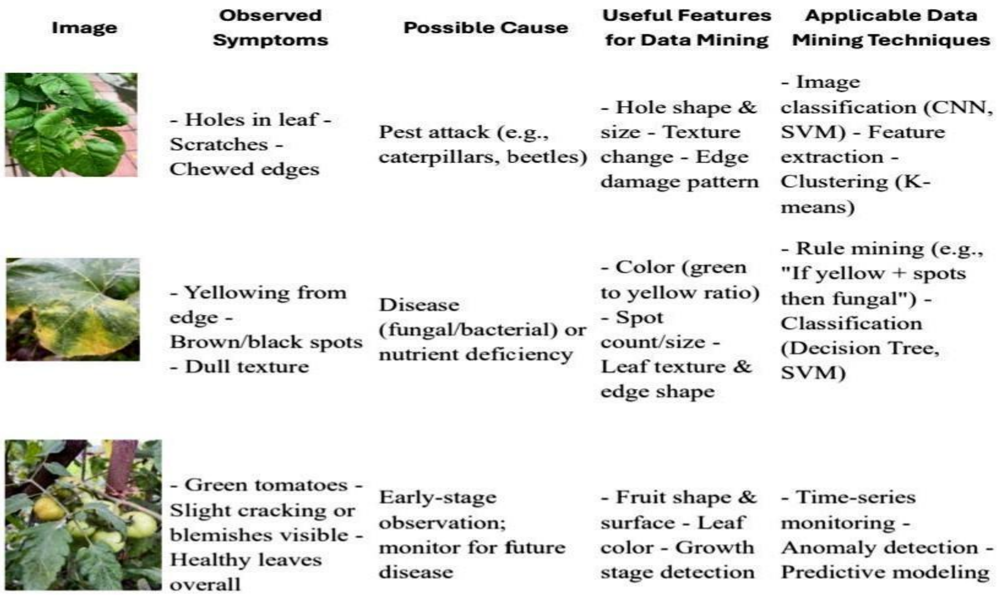

Image	Observed Symptoms	Possible Cause	Useful Features for Data Mining	Applicable Data Mining Techniques
	- Holes in leaf - Scratches - Chewed edges	Pest attack (e.g., caterpillars, beetles)	- Hole shape & size - Texture change - Edge damage pattern	- Image classification (CNN, SVM) - Feature extraction - Clustering (K-means)
	- Yellowing from edge - Brown/black spots - Dull texture	Disease (fungal/bacterial) or nutrient deficiency	- Color (green to yellow ratio) - Spot count/size - Leaf texture & edge shape	- Rule mining (e.g., "If yellow + spots then fungal") - Classification (Decision Tree, SVM)
	- Green tomatoes - Slight cracking or blemishes visible - Healthy leaves overall	Early-stage observation; monitor for future disease	- Fruit shape & surface - Leaf color - Growth stage detection	- Time-series monitoring - Anomaly detection - Predictive modeling

7. FUTURE SCOPE

1. Creation of Unified Agricultural Databases:
Future work should focus on developing open, large-scale, and crop-specific multimodal datasets combining sensor data, weather parameters, soil conditions, and UAV imagery. Standardization will enable fair comparison and model benchmarking across regions.

2. Explainable and Interpretable AI Models:
As farmers and agronomists rely on automated predictions, it is crucial to build explainable AI systems that can justify their predictions in human-understandable terms—showing which environmental variables or image features led to a pest/disease alert.

3. Integration with Real-Time IoT and Edge Computing:
Future agricultural systems can leverage low-cost IoT devices and edge AI computing to process data directly in the field. This will reduce latency and enable real-time pest outbreak alerts even in low-connectivity regions.

4. Cross-Regional Transfer Learning:
Transfer learning and domain adaptation can make models trained in one geographical region applicable to other climatic zones, addressing the generalization problem of AI in agriculture.

5. Causal and Mechanistic Modelling:
Combining data-driven ML models with biological and ecological process models will enable deeper insights into pest life cycles, reproduction patterns, and the impact of environmental interventions.

6. Sustainability and Climate Change Impact Studies:
With increasing climate variability, integrating predictive analytics into climate-resilient pest management strategies will help policymakers and farmers anticipate shifts in pest populations and emerging diseases.

8. CONCLUSION

The integration of data mining with modern artificial intelligence techniques has transformed pest and disease prediction into a proactive and data-driven process. From the comprehensive review of thirty research papers, it is evident that no single model suffices across all crops or regions; instead, hybrid and ensemble approaches achieve the best performance. Plant disease and pest detection approaches based on deep learning combined edge detection and feature extraction have wide advancing projections and high potential, in contrast to standard image processing techniques, which handle these jobs in various phases and linkages. [3]This survey presented an insight into existing research addressing the application of ML-based techniques for forecasting, detection, and classification of diseases and pests.[4]

Deep learning-based models, particularly CNNs, have shown exceptional ability in visual disease detection from leaf and UAV imagery, whereas hybrid ARIMA–LSTM and GRU-based models effectively forecast temporal pest occurrences using climatic and biological data. Additionally, GAN-based synthetic data generation addresses the critical problem of dataset imbalance, leading to improved accuracy and robustness in outbreak prediction.

The article provides a comprehensive overview of the current advancements in the technology for detection and identification of plant diseases used in agriculture. The initial goal was to assess recent plant disease detection and identification technologies that were employed with a focus to use remote sensing, ML and DL algorithms.[22]Agriculture is suffering from a number of problems; plant diseases and pests are contributing as the most devastating factor.

Diseases on the leaves of tomato plants have a negative impact on both quality and yield. Deep learning has shown great potential in improving tomato disease and pest detection, offering high accuracy and efficiency compared to traditional methods.[33]

Despite technological progress, real-world application remains limited due to issues such as data heterogeneity, lack of standardized datasets, and regional variations in pest behavior. The synthesis reveals that multimodal data fusion (combining imagery, IoT sensor streams, and historical trends) provides the most resilient and scalable solutions for precision agriculture. Hence, integrating these models into real-time decision-support systems (DSS) is a key step toward sustainable and intelligent crop protection.

9. DOCUMENTATION

The project documentation provides a comprehensive description of each phase of the data- mining-based pest and disease prediction system. It follows the standard structure proposed in agricultural ML studies such as [31] Liu & [2] Wang (2021) and [35] Mittal et al. (2024), which emphasize the importance of data preprocessing, model training, and evaluation pipelines for agricultural prediction systems. According to [32] Kang et al. (2023), accurate pest detection requires modular documentation covering system design, input data flow, and result interpretation to ensure reproducibility and field usability.

10. REFERENCES

[1] Shoaib, M. et al., “An Advanced Deep Learning Models-Based Plant Disease Detection Survey,” Applied Sciences, 2023.

[2] Wang, M., “Hybrid ARIMA–LSTM Model for Pest and Disease Forecasting in Sugarcane,” MDPI Agriculture, 2025.

[3] Upadhyay, A., “Deep Learning and Computer Vision in Plant Disease Detection,” Computers in Agriculture, 2025.

[4] Domingues, T., “Machine Learning for Detection and Prediction of Crop Diseases,” MDPI Sensors, 2022.

[5] Krishna, M.S., “Plant Leaf Disease Detection Using Deep Learning,” MDPI Plants, 2025.

[6] He, T., “Deep Learning-based Time Series Prediction for Precision Crop Protection,” Frontiers in Plant Science, 2025.

[7] Ibrahim, E.A., “An Expert System for Insect Pest Population Dynamics,” Computers and Electronics in Agriculture, 2022.

[8] Matese, A., “UAV-based Hyperspectral Imaging for Crop Monitoring,” Remote Sensing, 2024.

[9] Ferentinos, K.P., “Deep Learning Models for Plant Disease Detection and Diagnosis,” Computers and Electronics in Agriculture, 2018.

[10] Ali, A.H., “Ensemble Deep Learning Architectures for Plant Disease Classification,” Journal of Big Data, 2024.

[11] Pandit, P., “Hybrid Time Series Models with Exogenous Variables,” Applied Artificial Intelligence, 2023.

[12] Palma, G.R., “Pattern-based Prediction of Population Outbreaks,” Agricultural Informatics Journal, 2023.

[13] Tai, C.Y., “Using TimeGAN to Synthesize Sensing Data for Pest Incidence Forecasting,” Sustainability, 2023.

[14] Shahi, T.B., “Advances in Crop Disease Detection Using UAV Remote Sensing,” Remote Sensing Letters, 2023.

[15] García-Vera, Y.E., “Hyperspectral Image Analysis and Machine Learning in Crops,” MDPI Remote Sensing, 2024.

[16] Zhu, H., “Intelligent Agriculture: Deep Learning in UAV-based Remote Sensing,” Frontiers in Plant Science, 2024.

[17] Elfouly, M.K., “Deep Learning-based Large-scale Plant Disease Framework,” Journal of Big Data, 2025.

[18] Wahyono, T., “Enhanced LSTM Forecasting for Crop Pest Attacks,” ICICEL, 2020.

[19] Paul, B., “Advancements in AI-based Pest and Disease Detection in Agriculture: A Comprehensive Review,” SSRN Preprint, 2024.

[20] Nautiyal, M., “Revolutionizing Agriculture: A Review on AI Tools,” Agricultural Systems, 2025.

[21] Xu, M., “Plant Disease Recognition Datasets in the Age of Deep Learning,” Computers in Agriculture, 2024.

[22] Khan, S.U., “A Review on Automated Plant Disease Detection,” Plant Science Journal, 2025.

[23] ResearchGate, “AI-driven Insect Detection and Real-time Monitoring in Greenhouses,” ResearchGate Preprint, 2025.

[24] ArXiv, “LeafGAN: Data Augmentation for Plant Disease Diagnosis,” arXiv preprint, 2020.

[25] ArXiv, “CropDocNet for Potato Late Blight Detection from UAV Hyperspectral Imagery,” arXiv preprint, 2021.

[26] ResearchGate, “Review on Detection and Prediction of Crop Disease Using Machine Learning,” ResearchGate Review, 2024.

[27] García-Vera, Y.E., “Hyperspectral Applications in Crop Disease Detection,” MDPI, 2024.

[28] Sciencedirect, “A Review on Machine Learning and Deep Learning for Plant Disease Detection,” ScienceDirect Review, 2024.

[29] ICICEL, “Bidirectional LSTM for Crop Pest Forecasting,” ICICEL Conference Proceedings, 2020.

[30] ResearchGate, “Remote Sensing Using UAVs for Detecting Crop Diseases,” ResearchGate Survey, 2023.Udshkt

[31] Liu J, Wang X. Plant diseases and pests detection based on deep learning: a review. Plant Methods. 2021;17:22. Doi:10.1186/s13007-021-00722-9.

[32] Kang H, Ai L, Zhen Z, Lu B, Yi P, et al. A Novel Deep Learning Model for Accurate Pest Detection and Edge Computing Deployment. Insects (MDPI). 2023;14(7):660. Doi:10.3390/insects14070660.

[33] Jelali M, et al. Deep learning networks-based tomato disease and pest detection: a first review of research studies using real field datasets. Frontiers in Plant Science. 2024.

[34] Lee S, Yun CM. A deep learning model for predicting risks of crop pests and diseases from sequential environmental data. Plant Methods. 2023. Doi:10.1186/s13007-023- 01122-x.

[35] Mittal M, Gupta V, Aamash M, Upadhyay T. Machine learning for pest detection and infestation prediction: A comprehensive review. WIREs Data Mining and Knowledge Discovery. 2024.

[36] Xuanyu, C. et al. (2024). Time series prediction of insect pests in tea gardens using DL models. Journal of the Science of Food and Agriculture.

[37] Boopathi, T. C Singh, S. (2015). Forecasting of Lychee pest incidence using ARIMA time- series. Journal of Insect Science.

[38] Gonçalves, J. et al. (2022). Edge-Compatible Deep Learning Models for Detection of Pest Outbreaks in Viticulture. Agronomy.