

Watermelon Fruit and Leaf Disease Prediction using VGG16

Soukhya Madan Nayak¹, Sharanamma Katti¹, Prema T. Akkasaligar¹, Abhishek Angadi¹, and Prem V¹

Department of Computer Science and Engineering, KLE Tech. University's Dr. MSSCET, Belagavi, India
premaakkasaligar.mss@kletech.ac.in

Abstract. Watermelon leaf diseases, such as mosaic virus and downy mildew, fruit disease such as anthracnose, pose a significant threat to crop yield and quality. Traditional manual diagnosis methods are time-consuming and prone to inaccuracies. It requires the development of automated systems that leverage machine learning and deep learning to achieve faster and more reliable identification. The study focuses on the classification of watermelon diseases into four categories: healthy, mosaic virus, anthracnose, and downy mildew. The primary challenge lies in developing a robust automated system capable of handling the complexities of disease identification with high accuracy from leaf and fruit images. To address this, a dataset is collected from the Regional Horticulture Research Station, Patuakhali, Bangladesh, characterized by its diversity and complexity. The proposed solution employs a deep learning-based VGG16 model, which has demonstrated its effectiveness in classifying watermelon diseases with enhanced precision. It offers a practical and scalable approach to disease diagnosis in agricultural systems.

Keywords: Watermelon leaf disease, VGG16, Anthracnose, Downy Mildew, Mosaic Virus

1 Introduction

Watermelon diseases are caused by the interaction of pathogens in environmental conditions, and agricultural practices. Viruses, fungi, bacteria, and oomycetes are some of the pathogens that infect watermelon crops. Examples of common diseases include Mosaic Virus, anthracnose, and downy mildew. These diseases are mainly transmitted through insects, contaminated tools, and water. Environmental factors such as high humidity, temperature fluctuations, and water stress contribute to the proliferation of these pathogens, while improper agricultural practices such as overcrowding, inadequate crop rotation, and poor pest management further increase susceptibility. To predict these diseases, modern approaches such as machine learning, computer vision, and IoT sensors are increasingly being used.

Watermelon crops, a crucial agricultural commodity, are highly susceptible to a variety of fruit disease like anthracnose and leaf disease like mosaic virus, downy mildew. These diseases significantly impact both the quality and yield of crops, undermining agricultural productivity and threatening the economic stability of farmers. Timely and accurate disease detection is critical to mitigate crop losses and ensuring food security. However, the current reliance on manual observation methods is labor-intensive, subjective, and prone to inaccuracies, highlights the urgent need for automated, efficient solutions.

The increase in global demand for food security, exacerbated by the effects of climate change on the prevalence of plant diseases, requires the development of innovative methods for disease detection. Traditional approaches that depend on human observation often result in a delayed diagnosis, leading to substantial economic losses. Automating the detection process not only enhances efficiency but also ensures consistency and accuracy, making it an essential step toward sustainable agricultural practices.

Watermelon fruit and leaf diseases significantly impact crop yield and quality, with anthracnose, downy mildew, and mosaic virus being the most prevalent. Anthracnose, caused by the fungus *colletotrichum lagenarium*, is responsible for about 35% of infections in humid regions. It manifests as yellow or brown leaf spots that can lead to defoliation and reduce yields by up to 30%. Downy mildew, triggered by the pathogen *pseudoperonospora cubensis*, affects approximately 25% of plants in wet climates, causing angular yellow spots on leaves that eventually turn necrotic. Mosaic virus, commonly spread by aphids, results in a patchy yellow and green pattern on leaves and stunted growth, with about 15% of plants showing symptoms. Healthy watermelon plants, free from these diseases, comprise around 25% of the surveyed population, often attributed to effective disease management practices. To avoid these diseases, emphasize the need for integrated pest management and resistant cultivars to mitigate losses [1].

Numerous methods for detecting plant diseases have been made possible by recent developments in computer vision and machine learning. With different degrees of effectiveness, traditional machine learning models like random forests (RF) and support vector machines (SVM) have been used. However, these approaches frequently call for a great deal of feature engineering, which takes a lot of time and is highly dependent on domain knowledge. Convolutional neural networks (CNN), have been the cutting-edge method for image-based problems in recent years. They are especially useful for disease detection based on visual data because of their ability to automatically extract hierarchical features from raw images.

Despite their advantages, existing methodologies face notable limitations. Classical machine learning models such as SVM and RF are heavily based on handcrafted features and struggle to generalize across diverse datasets. Although CNN eliminates the need for manual feature extraction, their performance is often hindered by the limited availability of large, labeled datasets, which are critical for training effective models. In addition, CNN are computationally intensive, posing challenges for deployment in resource-constrained environments. These limitations underscore the need for a solution that balances accuracy, scalability, and computational efficiency.

The development of an effective disease detection system for watermelon crops presents additional challenges. One of the most pressing issues is the scarcity of labeled data, which restricts the creation of highly accurate models. Class imbalance, where certain diseases are underrepresented in the dataset, leads to biased model predictions. Furthermore, realworld variability in images, including differences in lighting, camera angles, and background, adds complexity to the detection process. The proposed system is scalable and adaptable to various diseases and types of crops, ensuring broad applicability and sustainability.

The present paper proposes a CNN-based approach to address these challenges in the detection of watermelon fruit and leaf diseases. Using CNN, the system automatically learns the features from images, eliminating the need for manual extraction. The problem of limited labeled data is addressed through data augmentation techniques, which artificially expand the data set and improve the robustness of the model. The proposed system is designed to classify diseases such as mosaic virus, anthracnose, and downy mildew with high accuracy while maintaining computational efficiency. Moreover, the design is scalable and extensible, able to generalize other types of crops and diseases with minimal modifications. The proposed model also incorporates insights from prior studies, such as Dinesh Arora's hybrid RNN-RF model for detecting watermelon mosaic virus. While this model achieved an impressive accuracy of 99.08% using 16,000 images, it faced challenges such as long training times and limited data diversity [2]. The proposed CNN-based approach addresses these limitations, offering a more efficient and versatile solution for disease detection.

The layout of the paper is organized as follows: the introduction (section 1), literature review (section 2), proposed methodology (section 3) along with results and analysis (section 4) and conclusion (section 5).

2 Literature survey

Watermelon is a vital crop with high economic and nutritional value, but its cultivation faces challenges from various diseases that can significantly reduce yield. Hence, machine learning (ML) and deep learning (DL) techniques have emerged as effective tools for disease detection and classification. Studies have explored methods such as SVM, RF, KNN, ANN, CNN, and hybrid models, demonstrating their potential for accurate diagnosis. However, recurring challenges such as high computational costs, limited data diversity, and restricted generalization ability highlights the need for robust, scalable, and efficient systems to enhance disease detection and management.

In [3], a review on plant leaf diseases detection and classification employed for machine learning methodologies, including SVM, RF, KNN, and ANN. The limitations of the work includes the computational cost for SVM and the suggestion for adopting hybrid models. The identified gaps are the need for higher efficiency in classification and broader applicability to diverse datasets. In [4], a cognitive system for watermelon disease diagnosis is proposed using a cognitive expert system (CLIPS and Delphi XE10.2). Despite of providing a structured diagnostic framework, the study is limited to seven specific diseases, highlighting the gap for extending the system to a wider range of diseases for better scalability. In [5], researchers implemented RGB color extraction for disease detection in watermelon leaves. The approach faced challenges related to image quality dependency, with a key gap identified being the need for higher-resolution camera models for better detection and classification outcomes. In [6], the detection of watermelon mildew is studied using hyperspectral images combined with machine learning techniques such as MLP and DT. The limitations of the

study includes its low applicability to earlystage disease detection, with the need for further improvements in early detection methodologies highlights a significant gap. In [7], a study focuses on early detection of plant diseases using CNN and SVM on tomato leaves. While effective for tomato crops, the study faced limitations in generalization to other crops, indicating a need for a more versatile detection model capable of addressing a wider range of plants and diseases. In [8], a KNN algorithm is used for the detection of watermelon and cucumber diseases. Although the approach is effective, it has limitations related to high computational cost and restricted generalizability, with gaps identified in adapting the system for diverse and complex datasets. In [9], a hybrid RNN and RF model is developed for detecting watermelon mosaic virus. This model faced challenges such as long training times and limited diversity in image datasets, with the identified gaps including the need for models that are more efficient and capable of handling greater variability in data. In [10], CNN-based detection and classification of plant leaf diseases are explored using multiple datasets. The limitations are related to variations in data performance and issues in generalizing to uncontrolled environments. This highlights a significant gap in developing robust system adaptable to real-world conditions. In [11], a stacked RNN approach combined with image processing techniques is employed for watermelon disease detection. This work is limited by its high dependency on image quality, and the gap identified is the need for better segmentation techniques to improve model performance.

The Table 1 Shows literature review of watermelon disease detection techniques for the year 2024 These studies collectively emphasize the growing importance of advanced computational techniques like CNN and hybrid models while highlighting recurring challenges such as limited data diversity, high computational costs, and restricted generalizability.

Table 1: Literature Review of Watermelon Disease Detection Techniques for the year 2024

Title	Methodology	Dataset Description	Performance Parameters	Gaps Identified
Hybrid RNN and RF for WMV severity [2]	RNN and Random Forest	16000 images of WMV	Accuracy: 99.08%	Data variety and image quality issues
UAV data for melon leaf disease detection [11]	UAV, YOLOv8 for detection	UAV images	Accuracy: 83.2%	Need for larger-scale validation
Detection of bacterial infection in melons [12]	Imaging data, ML (ANN, SVM)	Melon leaves dataset	ANN accuracy: 99.1%	Further refinement needed
Federated learning for watermelon disease detection [13]	Federated Learning (FL), CNN	Five distinct disease data subsets	Accuracy: 97%	Improve client communication reliability

3 Proposed methodology

Watermelon diseases, like mosaic virus and anthracnose significantly cause crop damage. Automated systems based on machine learning and deep learning have the capability to attain faster and more accurate diagnosis as compared to the traditional manual methods. The identification of diseases of watermelon as namely healthy, mosaic virus, anthracnose, and downy mildew. The challenge is to design an automated system that classifies watermelon diseases with maximum accuracy from the images of the leaves.

The primary objectives of the present paper are:

- To enable the early detection of watermelon diseases.
- To minimize crop losses through timely intervention.
- To achieve high classification accuracy by utilizing deep learning models.
- To distinguish between different disease categories effectively.

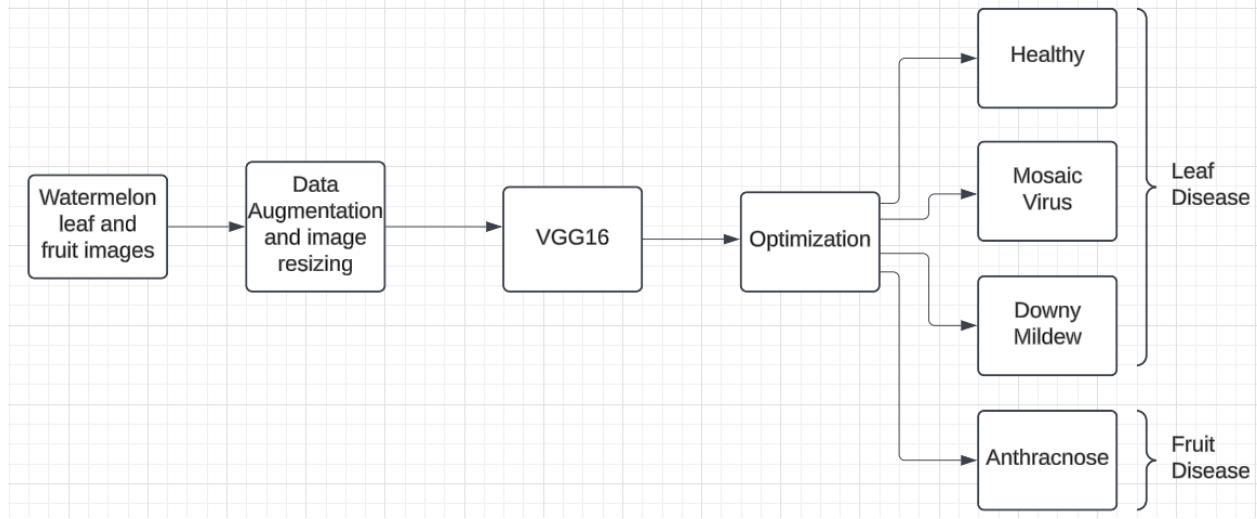


Fig. 1: Proposed methodology to classify watermelon fruit and leaf disease

The proposed methodology for classifying watermelon fruit and leaf disease is illustrated in Fig. 1. It shows the essential modules to ensure precise and effective classification. The main goal of the data gathering phase is to capture images of both healthy and diseased watermelon leaves from the regional horticulture research station in Patuakhali, Bangladesh. The dataset contains various categories, including downy mildew, anthracnose, mosaic virus, and healthy.

In the data preprocessing stage, to increase robustness and avoid overfitting, data augmentation methods such as rotation, flipping, zooming, and shifting are applied. This process produces a large dataset comprising both enhanced and original images. Images are standardized to a uniform size (e.g., 224x224 pixels) and normalized to scale pixel values between -1 and 1 for better convergence during training. The dataset is then split into training (70%), validation (15%), and testing (15%) subsets, ensuring balanced representation of all categories across the splits.

The model has well-known CNN architecture VGG16. The architecture is fine-tuned by customizing convolutional layers, activation functions (ReLU), pooling layers, and fully connected layers to suit the dataset's complexity. The training process involves monitoring validation performance to prevent overfitting while optimizing the model using techniques such as Adam optimizers, learning rate scheduling, dropout, batch normalization, and early stopping.

Finally, the evaluation is made to the trained model. It is tested on the reserved test set, and its performance is assessed using metrics like accuracy, precision, recall, F1-score, and a confusion matrix. Error analysis of misclassified images is conducted to identify potential improvements. The systematic approach ensures high accuracy and reliability in detecting watermelon diseases. The detailed steps of implementation are represented using Algorithm 1

Algorithm 1: Watermelon Disease Classification

Input: Image dataset of watermelon fruit and leaf.

Output: Classification of images into healthy, mosaic virus, anthracnose, and downy mildew categories.

Start

Step 1: Preprocess the images: resize images to pixels and normalize pixel values between -1 and 1.

Step 2: Perform data augmentation for training by applying transformations such as rotation, width/height shift, zoom, shear, and horizontal flip to increase dataset variety and improve model robustness.

Step 3: Load the pre-trained VGG16 model, excluding the top layer, and load pre-trained ImageNet weights for feature extraction.

Step 4: Freeze the layers of the base model to retain pre-trained features by setting all layers of the VGG16 base model to non-trainable.

- Step 5:** Build the model by adding global average pooling, a fully connected layer with ReLU activation, a dropout layer, and a softmax output layer for classification.
- Step 6:** Compile the model using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric.
- Step 7:** Define callbacks to implement early stopping, learning rate reduction, and model checkpointing to monitor training and save the best model.
- Step 8:** Train the model on the training set and validate it on the validation set, applying the defined callbacks.
- Step 9:** Save the trained model to disk after training completes.
- Step 10:** Plot training and validation accuracy and loss over epochs to monitor model performance.
- Step 11:** Evaluate the model on the test dataset to measure its performance and generalization ability.
- Step 12:** The trained model is used to predict the disease category with the highest probability.
- Step 13:** Output the result: Display the predicted disease type based on the highest predicted class.

End

4 Results and discussion

The machine used for the experiments has the following specifications: Operating System: Windows 11, Processor: AMD Ryzen 5 5500U with Radeon Graphics, RAM: 8 GB. The Python programming language is used to implement the model. The two main libraries for creating deep learning models are TensorFlow and Keras. Pandas and NumPy are used for effective data analysis and manipulation. The core of the suggested approach is the creation of sophisticated machine vision algorithms for the early detection of watermelon diseases. The dataset is specifically made to classify leaf disease mosaic virus, healthy, downy mildew and fruit disease anthracnose. The dataset consists of a total of 6930 images, including 1155 original images and 5775 augmented images. Each image in the dataset represents a detailed visual depiction of a watermelon leaf and fruit, providing clear insights into its texture, color, and overall health status[14].

In Fig. 2, mosaic virus displays characteristic leaf mottling and discoloration. The healthy leaf shows uniform color and texture without visible damage. Anthracnose includes dark, necrotic spots indicating fungal infection. Downy Mildew highlights yellowish, angular spots and a downy growth on the underside of leaves[14].



Fig. 2: Sample images: (a) Mosaic Virus, (b) Downy Mildew, (c) Healthy Leaf, (d) Anthracnose.

All images have been preprocessed to a resolution of 224x224 pixels for model input, ensuring uniformity and compatibility with state-of-the-art machine vision frameworks. Augmentation is performed using techniques such as rotation, zoom, width and height shifts, shearing, and horizontal flipping. Normalization to a pixel value of range -1 to 1 is performed. These steps enhance the model's robustness by creating diverse variations of the training data.

Pre-trained weights on the imageNet dataset are fed into the main VGG16 model. New fully connected layers are added, its convolutional layers are frozen to preserve learning feature extraction capabilities. The dimensionality of the retrieved features is decreased using a globalaveragepooling and then a dense layer with 256 units utilizing ReLU activation is applied. To reduce overfitting, dropout regularization is used at a rate of 0.5. Lastly, the input is categorized into one of the four groups by a dense output layer with softmax activation. The Adam optimizer is used to assemble the model, and a learning rate of 0.0005 is initially set. Accuracy is chosen as the evaluation metric, while categorical crossentropy is employed as the loss function.

The model is trained for a maximum of 30 epochs with early stopping to prevent overfitting. Training and validation metrics, including accuracy and loss, are recorded and analyzed. The accuracy improved steadily during the initial epochs before stabilizing. The validation loss followed a similar trend, indicating the model's ability to generalize effectively to unseen data. Plots of accuracy and loss against the number of epochs are shown in Figures 3 and 4 respectively. These visualizations highlights the points where early stopping is triggered, showcasing the model's stability.

After training, the model is evaluated on a separate test dataset to assess its performance. The final trained model is saved for future inference tasks. The combination of data augmentation, a robust pre-trained backbone of (VGG16), and regularization techniques contributed significantly to the model's success. These results validate the proposed approach for watermelon leaf and fruit disease, paving the way for its application in real-world agricultural monitoring systems.

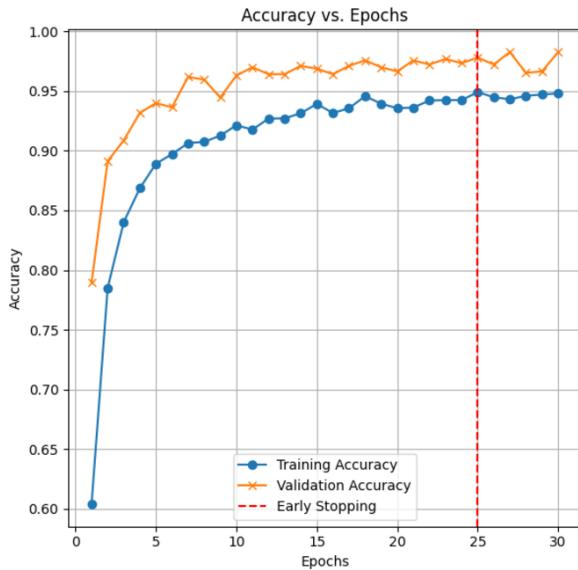


Fig. 3: Accuracy curve for VGG16

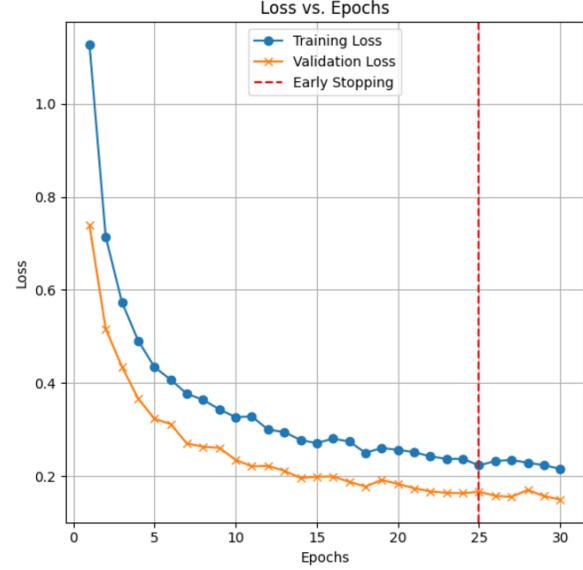


Fig. 4: Loss graph for VGG16

In Fig. 3, the graph shows that the model's accuracy increased steadily during training, but it started to overfit around epoch 15. Early stopping at epoch 25 prevented further overfitting. In Fig. 4, the model's training loss decreased steadily, while the validation loss decreased initially but plateaued around epoch 10. Early stopping at epoch 25 is appropriate as it prevented overfitting. The Fig. 5, represents confusion matrix. further, it is observed that, the model to perform best on the downy mildew class, with a high number of correct predictions (114) and relatively few misclassifications. The model struggles with the anthracnose class, having a significant number of false positives and false negatives. The mosaic virus class also shows some misclassifications, particularly with downy mildew.

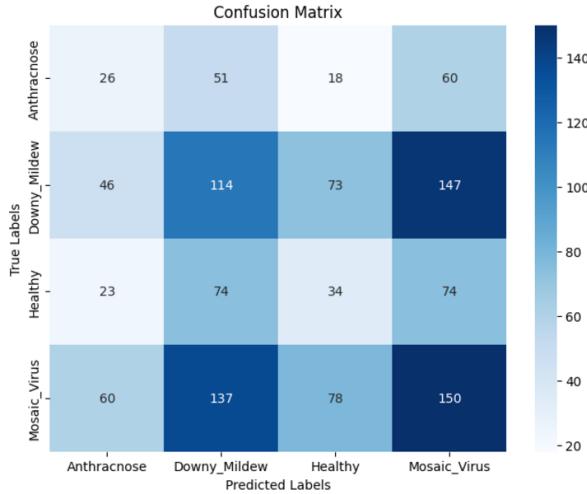


Fig. 5: Confusion matrix VGG16

The Table 2, shows that the model consistently achieves high precision across all classes, meaning that when it predicts a class, it is highly likely to be correct. The Recall demonstrates high value, indicating that it is able to identify most of the positive cases for each class. The F1-score balances precision and recall, is found to be high for all classes, suggesting a well-rounded performance. The final training accuracy is 94.81%, while the test accuracy reaches 98.8%, demonstrating effective generalization to unseen data. The classification report indicates high precision, recall, and F1-scores across all classes, with an overall accuracy of 98.79%. The results highlight the model's high performance in distinguishing between various watermelon leaf and fruit diseases, with minimal misclassification and high reliability. The Table 3, shows the performance comparison with existing model. It is observed that the, proposed model shows better performance as compared to Siti Nur Aisyah method across all evaluation matrices. Its robust generalization and precise classification make more suitable for the task.

Table 2: Classification performance of VGG16 model

Class	Precision	Recall	F1-Score
Anthracnose	100.0	100.0	100.0
Mosaic Virus	97.67	99.05	98.36
Healthy	99.5	98.53	99.01
Downy Mildew	99.2	98.15	98.67
Accuracy		98.79	
Macro Average	98.8	98.79	98.79

Table 3: Comparision of Weighted Average Performance Between VGG16 and Siti Nur Aisyah Models

Metric	VGG16	Siti Nur Aisyah [11]
Accuracy(%)	98.79	83.0
Recall	98.79	-
Precision	98.8	-
F1 score	98.79	-

5 Conclusion

Watermelon diseases occurs due to various factors such as fungal infections (e.g., anthracnose and downy mildew), viral infections (e.g., mosaic virus), bacterial infections, and environmental conditions like excessive moisture or poor soil health. These diseases not only affect the health and yield of the crop but also have economic consequences for farmers. The present paper proposes watermelon leaf and fruit disease prediction using VGG16. The use of data augmentation and regularization techniques like dropout minimized overfitting, ensuring strong generalization to unseen data. The inclusion of custom classification layers tailored to the dataset ensured that the model adapted well to the specific task, achieving an overall accuracy of 98.79values across disease categories demonstrate the robustness and reliability of the model in distinguishing between healthy, mosaic virus and downy mildew leaf disease, anthracnose fruit disease. By leveraging the pre-trained VGG16 model, the approach benefits from transfer learning, effectively extracting complex features from the leaf images while reducing the computational resources and time required for training from scratch. In summary, the VGG16-based model is well-suited for watermelon disease classification due to its ability to generalize effectively, its efficiency in leveraging transfer learning, and its reliable performance across all metrics. This makes it a valuable tool for real-world agricultural applications, offering a reliable method to predict and manage watermelon diseases based on visual symptoms. By integrating such predictive systems into agricultural practices, farmers can make timely decisions to protect their crops, ensuring better yields and reducing losses.

References

1. Moya-Ruiz C, Gómez P, Juárez M. Occurrence, Distribution, and Management of Aphid-Transmitted Viruses in Cucurbits in Spain. *Pathogens*. 2023 Mar 7;12(3):422. doi: 10.3390/pathogens12030422. PMID: 36986344; PMCID: PMC10057868.
2. D. Arora, K. Mehta, A. Kumar, and S. Lamba, “Evaluating Watermelon Mosaic Virus Seriousness with Hybrid RNN and Random Forest Model: A Five-Degree Approach,” in Proceedings of the 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICRITO61523.2024.10522240.
3. S. Swain, S. K. Nayak, S. Sucharita, and A. Barik, “A Review on Plant Leaf Diseases Detection and Classification Based on Machine Learning Models,” 2020. [Online]. Available: <https://www.semanticscholar.org/paper/A-Review-on-Plant-Leaf-Diseases-DetectionandBased-Swain-Nayak/5ac88f43e168d1fb40d40aae9186470bae3594d7>.
4. B. S. Abu-Nasser and S. S. Abu-Naser, “Cognitive Expert System for Watermelon Disease Diagnosis and Treatment,” *International Journal of Academic Information Systems Research*, vol. 2, no. 7, pp. 1-7, 2018, hal-01855441.
5. S. B. Kutty , “RGB Color Extraction-Based System for Classifying Watermelon Leaf Diseases,” in Proceedings of the 2021 Business Engineering and Industrial Applications Colloquium (BEIAC), 2021, doi: 10.1109/BEIAC.2021.6560170. [Online]. Available: <https://www.researchgate.net/publication/261264177>
6. J. Abdulridha, Y. Ampatzidis, J. Qureshi, and P. Roberts, “Identification and Classification of Downy Mildew Severity Stages in Watermelon Utilizing Aerial and Ground Remote Sensing and Machine Learning,” *Frontiers in Plant Science*, vol. 13, p. 791018, 2022, doi: 10.3389/fpls.2022.791018.
7. S. S. Harakannanavar , “CNN-Based Early Detection of Plant Diseases in Tomato Leaves,” *Global Transitions Proceedings*, vol. 3, no. 11, pp. 1-8, 2022, doi: 10.1016/j.gltip.2022.03.016. [Online]. Available: <https://www.researchgate.net/publication/359690037>.

8. J. Doe , “Detection of Cucumber and Watermelon Diseases Based on Image Processing Techniques Using K-Means Algorithm,” International Journal of Multidisciplinary Research and Growth Evaluation, vol. 4, pp. 38-46, 2023, doi: 10.54660/.IJMRGE.2023.4.6.38-46.
9. H. S. Jayswal , “Various Plant Disease Detection Methods (CNN, SVM, Mobile Apps),” Plant Leaf Disease Detection and Classification using Conventional Machine Learning and Deep Learning, Jan. 2020. [Online]. Available: <https://www.researchgate.net/publication/352787124>
10. D. Jayakumar , “Automatic Prediction and Classification of Diseases in Melons Using Stacked RNN Based Deep Learning Model,” in Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262414.
11. S. N. A. Robi , “Utilizing UAV Data for Neural Network-Based Classification of Melon Leaf Diseases in Smart Agriculture,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.01501119.
12. M. Pineda , “Detection of Bacterial Infection in Melon Plants by Classification Methods Based on Imaging Data,” Frontiers in Plant Science, vol. 9, p. 164, 2018, doi: 10.3389/fpls.2018.00164.
13. A. Suryavanshi , “Agriculture Farming Evolution: Federated Learning CNNs in Combatting Watermelon Leaf Diseases,” in Proceedings of the 2024 Asia Pacific Conference on Innovation in Technology (APCIT), Mysore, India, 2024, pp. 1-6, doi: 10.1109/APCIT62007.2024.10673452.
14. <https://www.kaggle.com/datasets/sujaykapadnis/watermelon-disease-recognition-dataset>