Article

Machine Learning and Plant Disease Detection: A Brief Review

Swarajya Lakshmi .P

Research Scholar, Department of Computer Science and Engineering
Manas Global University, Bhopal, Madhya Pradesh, India
Asst.Professor, Department of Computer Science and Engineering,
Anurag Universiy, Hyderabad, Telangana, India

Citation:
Swarajya lakshmi. P.

Machine Learning and Plant Disease detection: A Review,
J.Adv.Sci.Eng&Tech,
1,1,2025, 1

Received on
17th July 2024

Revised on
11th September 2024

Accepted on
23rd December 2024

Abstract

The global food security and agricultural sustainability face a major threat from plant diseases. Effective management requires both early and accurate identification of these diseases. The development of Artificial Intelligence (AI) has led Machine Learning (ML) techniques to become promising tools for automated disease detection and classification. The paper presents an extensive evaluation of current progress in employing ML methods for plant disease detection. The role of image processing combined with deep learning and traditional ML classifiers in disease identification is discussed in detail. A detailed literature survey is provided followed by a method- dataset-performance metrics comparison. The paper also discusses future research directions and challenges.

Key words: Machine learning, Plant disease detection, Deep learning, Image processing

1. Introduction

Agriculture serves as the base of the worldwide food supply and functions as a vital component for many national economies. The spread of plant diseases leads to significant reductions in crop yields along with decreases in quality. Traditional disease detection methods depend on trained pathologists as well as agronomists who perform manual visual assessments but these methods are labor-intensive and time-consuming while also prone to human errors. The integration of Machine Learning (ML) into agricultural practices brings about a transformative method for disease detection and classification. Intelligent systems developed through ML allow for the analysis of extensive image data sets to detect disease symptoms with high accuracy levels. Such systems can be integrated into mobile or drone-based systems for maximum scalability to support real-time applications in precision agriculture. The research investigates existing ML applications in plant disease detection using both traditional and deep learning methods as analysis bases. The study examines the datasets that researchers typically use as well as preprocessing methods and feature extraction techniques and performance assessment metrics in plant disease detection research.

2. Machine Learning in Plant Disease Detection

Machine learning encompasses a broad range of algorithms that enable computers to learn from data and make predictions or decisions without explicit programming. In the context of plant disease detection, ML models are typically trained on annotated images of plant leaves to identify patterns associated with various diseases. For instance, the use of drones has become increasingly common in recent years for plant disease detection, a trend largely spearheaded by the growth of drone technology.

A. Collecting Plant Disease Data

The foundational phase in developing a machine learning (ML) system for identifying plant diseases is acquiring relevant data. The system's accuracy and reliability heavily rely on the size, quality, and variety of the dataset used for training. Capturing high-quality images of leaves—both healthy and diseased—from diverse agricultural regions is vital. Data can be collected using a range of devices, including high-resolution DSLR cameras, portable smartphones ideal for field use, and drones that offer aerial coverage for large-scale crop monitoring.

Acquiring images directly from the field brings valuable natural variation into the dataset. This includes changes in light, background interference, overlapping leaves, and environmental elements like dirt or moisture. In contrast, collecting images in a controlled environment ensures uniform lighting and backgrounds, which may enhance lab-based model accuracy but often fails to perform well in real-world scenarios.

Widely used public datasets such as PlantVillage have significantly contributed to the field, offering over 50,000 annotated images across 14 crops and 26 diseases. To improve the model's adaptability, it's crucial to gather data from multiple regions, plant species, and different disease stages. Such diversity helps the model develop generalizable and robust patterns.

Accurate labeling of data is another key factor. While expert annotation by plant pathologists ensures high fidelity, it can be time-intensive and costly. To reduce these barriers, methods such as crowd annotation and semi-supervised learning are gaining popularity. Labels should ideally include disease name, intensity, and metadata such as time of year, geographic location, and plant variety. These data layers support the development of multi-dimensional models beyond simple disease classification.

Image clarity is vital for identifying fine details like spots, discoloration, or mold growth. While high-resolution images provide rich features, they demand more storage and computing power. Advanced imaging tools—such as thermal, multispectral, and hyperspectral cameras—can reveal plant stress markers invisible to the naked eye, often before symptoms are visible.

To improve the dataset further, artificial methods like rotation, zoom, and lighting adjustments simulate environmental changes, helping models become more adaptable. A frequent issue is data imbalance, where healthy samples dominate the dataset. Approaches like synthetic data creation via GANs (Generative Adversarial Networks), over/undersampling, or data balancing algorithms can help correct this.

Though satellites offer wide coverage, their resolution limits their utility for detecting diseases at the leaf level. Modern innovations such as IoT-enabled greenhouses and automated camera systems enable real-time image collection. These systems often integrate environmental sensors, capturing data on factors like soil moisture, air temperature, and humidity, which can enhance ML training by correlating conditions with disease symptoms.

Ethical data collection is essential—when working on private farms, informed consent and anonymization protocols should be followed. It's also critical to eliminate low-quality images (e.g., blurry or poorly lit) before training. Capturing metadata like GPS coordinates, crop variety, and date strengthens the value of datasets for future use.

Promoting open-access datasets can accelerate innovation but requires standardized formats and documentation to ensure usability across regions and languages. The timing of data collection should align with crop development stages to ensure symptoms are recorded accurately. Tools like LabelImg or VGG Image Annotator simplify labeling tasks, and collaboration with academic institutions can provide access to expert-verified data and real-world scenarios.

Synthetic data generation using techniques like GANs complements real-world data and helps simulate rare cases. Preprocessing steps such as image normalization and pixel value scaling ensure consistency across devices. Federated learning allows multiple stakeholders to train models collaboratively while maintaining data privacy. Ultimately, successful plant disease detection systems rely on a high-quality, varied, and well-labeled dataset as their foundation.

B. Image Preprocessing Techniques

Preprocessing prepares raw plant images for use in machine learning systems by standardizing input data and optimizing it for analysis. The first step typically involves resizing all images to a consistent dimension (e.g., 224×224 pixels), which ensures uniformity across the training set. Removing noise is equally important; images taken in natural settings may contain unwanted visual distortions. Filters like median, bilateral, or Gaussian are employed to reduce this noise.

Adjusting image contrast makes disease symptoms more visible. Methods such as histogram equalization and adaptive contrast enhancement are effective in highlighting features in underexposed images. Color normalization corrects discrepancies caused by different lighting setups and camera sensors, preserving consistent color patterns throughout the dataset.

Segmenting the leaf from its background is another vital task. Techniques like thresholding, K-means clustering, and contour analysis are used to extract the relevant portion of the image. Background clutter—such as other plants, soil, or sky—is removed so the focus remains on the infected areas.

Cropping the most significant part of the image, often where disease symptoms are concentrated, helps the model detect key features. Data augmentation methods like flipping and rotating can be applied at this stage to simulate orientation changes. Adjusting brightness and contrast further diversifies the training samples.

Changing the image's color space (e.g., converting from RGB to HSV or Lab) can emphasize color-specific disease indicators like yellowing or browning. Edge detection methods like the Canny or Sobel filters outline critical features such as lesions or fungal structures. Morphological operations, including dilation and erosion, refine boundaries and remove residual noise after segmentation.

When extracting numerical features such as texture or shape, scaling them through normalization or standardization ensures uniform data input. Color thresholding can pinpoint diseased areas when symptoms appear as unique hues or color bands. Techniques like Gaussian blur or sharpening filters are selectively used to either reduce visual clutter or enhance fine details.

Natural lighting often casts shadows; these must be removed to avoid misleading the model. Illumination correction techniques ensure lighting is uniform across the image. Preprocessing workflows are often automated using libraries like OpenCV or Scikit-Image to handle large datasets efficiently.

Batch normalization improves the consistency of input layers for neural networks, while dimensionality reduction via PCA (Principal Component Analysis) simplifies complex image features. ROI (Region of Interest) extraction ensures the model learns from the most relevant part of the leaf.

Adaptive preprocessing applies custom strategies depending on crop species, disease type, or image quality. Aligning leaves in a consistent direction supports more accurate pattern recognition. Filtering out duplicates or blurry images enhances dataset quality. GLCM (Gray-Level Co-occurrence Matrix) is frequently used for texture-based preprocessing.

AI-based preprocessing systems can intelligently choose the best processing techniques based on image characteristics. Algorithms that maintain color constancy help neutralize effects from varying illumination. Preprocessing for non-standard image formats like thermal or hyperspectral requires specialized steps such as noise filtering and dimensionality reduction.

For applications on mobile devices, lightweight preprocessing ensures that performance remains fast and efficient. Preprocessing reduces inconsistencies in the data, improves training accuracy, and enhances real-world performance of ML models. This step ensures that raw agricultural images are transformed into high-quality, analyzable inputs.

C. Extracting Informative Features

Feature extraction focuses on isolating the most relevant image characteristics that differentiate healthy and infected plants. These features serve as the input to machine learning algorithms. Color metrics like the average and standard deviation of RGB values often reveal early signs of disease through discoloration. Texture features—such as entropy, contrast, or smoothness—offer insight into changes in leaf surfaces.

Shape-based features capture geometric properties like the size and perimeter of lesions. Histograms of pixel intensity values summarize visual patterns. Edge detection can outline venation or necrotic patches, aiding diagnosis. Frequency domain tools like DFT (Discrete Fourier Transform) or DWT (Wavelet Transform) extract features that remain stable despite noise and lighting fluctuations.

Advanced filters like Gabor are especially useful for examining fine leaf textures. Keypoint detectors such as SIFT and SURF identify unique patterns that are resilient to scale or rotation changes. Dimensionality reduction techniques like PCA help condense information into a smaller set of powerful features. CNNs (Convolutional Neural Networks) automatically extract increasingly abstract features through their multiple layers—from basic edges to complex disease markers.

Pretrained CNN models, fine-tuned through transfer learning, allow feature extraction even with limited data. A combination of handcrafted and deep-learned features can enhance classification results. Feature selection strategies like Recursive Feature Elimination (RFE) and Information Gain refine the feature set, improving efficiency and performance.

Standardizing feature values is essential for consistent learning. Hyperspectral data introduces spectral features that detect stress responses invisible in standard RGB images. Time-based features from sequential images help monitor disease progression. Custom features may be crafted for specific crop-pathogen interactions.

Multi-scale approaches ensure both micro and macro features are captured. The BoVW (Bag of Visual Words) model groups local patterns into histograms for classification. LBP (Local Binary Patterns) is a popular texture descriptor used for subtle disease differentiation. For real-time systems, lightweight extraction methods prioritize speed without sacrificing accuracy.

Recent models employ attention mechanisms to prioritize crucial image regions during feature analysis. Hybrid models integrate empirical knowledge with data-driven insights for better interpretability. Automated pipelines reduce the reliance on domain experts, making systems more scalable. Feature fusion, where multiple feature types are combined, enhances the model’s flexibility across scenarios.

Resilient features are designed to perform well despite variation in lighting, background, or camera quality. Stable features ensure consistent performance across different datasets and environments. Accurate feature extraction ultimately reduces computational load and improves classification precision.

 |

D. Model Training

Training a model is a pivotal stage in crafting a machine learning-based system for identifying plant diseases. This phase involves teaching the algorithm to associate input data—typically features derived from leaf images—with the correct disease labels. Most implementations rely on supervised learning, where annotated image datasets guide the model's learning process.

Traditional algorithms like Support Vector Machines (SVM), Decision Trees, Random Forests, and k-Nearest Neighbors (k-NN) are commonly employed, especially when the features have been carefully engineered. These models tend to perform well when class distributions are balanced and feature spaces are clearly defined.

In recent years, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have gained popularity due to their ability to automatically learn hierarchical representations from image data. CNNs typically require a large volume of labeled images and use optimization strategies such as back propagation with loss functions (e.g., cross-entropy) to fine-tune model weights.

During training, the learning rate determines how quickly the model updates its parameters. Too high a rate can cause the model to overshoot optimal values, while too low a rate slows learning. Overfitting is a frequent concern, especially with limited data, and is managed using techniques like dropout (randomly deactivating neurons), early stopping (halting training before performance declines), and weight regularization.

To enhance generalization, training data is augmented by introducing variations—like flipped or rotated images—to mimic real-world variability. Cross-validation methods (e.g., k-fold) help evaluate the model's performance across different subsets of the dataset, ensuring reliability and minimizing bias.

Transfer learning is an effective approach where a model pretrained on large-scale datasets is adapted to the plant disease domain, significantly reducing training time and data requirements. Fine-tuning these models can yield state-of-the-art results even with limited data.

Advanced optimization algorithms such as Adam, RMSProp, or SGD with momentum improve convergence during training. Addressing class imbalance with strategies like weighted loss functions or resampling is critical for ensuring all disease types are learned equally.

In complex workflows, ensemble methods combine predictions from multiple models to boost overall accuracy. Cloud-based infrastructure offers scalable training capabilities, while on-device (edge) learning is emerging for privacy-focused and real-time applications. Federated learning allows collaborative model updates across decentralized data sources without transferring sensitive data.

Throughout training, key performance indicators like loss curves, accuracy metrics, and learning rates are logged for diagnostic purposes. Once the training phase is complete, the model is evaluated on an independent test set to assess its ability to generalize to unseen data. Ultimately, a well-trained model serves as the core engine of an automated, reliable plant disease detection solution.

E. Evaluating Model Effectiveness

After training, the next critical step is to measure how well the model performs in real-world scenarios. Evaluation provides objective criteria to determine whether the model accurately identifies plant diseases and can generalize beyond the training data. The most basic metric is accuracy, which calculates the percentage of correct predictions. However, this can be misleading in datasets where one class (e.g., healthy plants) is overrepresented. More nuanced metrics are required in such cases.

Precision indicates how many of the model's positive predictions are actually correct, while recall shows the proportion of true positives that were successfully detected. The F1-score offers a balanced metric by calculating the harmonic mean of precision and recall, especially useful when class distributions are uneven.

A confusion matrix provides a detailed view of classification results, showing the counts of true positives, false positives, true negatives, and false negatives. This helps diagnose specific errors—such as misclassifying one disease type as another.

The ROC curve (Receiver Operating Characteristic) and its Area Under the Curve (AUC) measure the model’s ability to distinguish between classes. A high AUC score indicates strong classification performance across different thresholds. Robust evaluation includes cross-validation, where the dataset is divided into several folds to ensure consistent performance across multiple runs. External validation using datasets collected under different conditions or from different regions tests the model’s ability to generalize.

In practice, real-time testing in agricultural fields is also essential. It reveals whether the model performs reliably in varying lighting, weather, and background conditions. For multi-class classification problems, macro-averaging treats all classes equally, while micro-averaging weighs them by frequency. Statistical significance testing, like t-tests or confidence intervals, helps determine whether observed improvements are meaningful rather than random variation.

Apart from accuracy, models are evaluated for efficiency—including prediction speed and memory usage—especially important for mobile or embedded applications. Model interpretability is another focus area. Techniques such as Grad-CAM, LIME, or SHAP help visualize which parts of an image influenced the model’s decision, fostering trust and transparency in its outputs.

Other dimensions of evaluation include robustness (resistance to noise or image distortions), fairness (equitable performance across crop types or regions), and scalability (ability to handle large datasets or deployments). Post-deployment, ongoing monitoring is vital to detect performance drift over time. Retraining may be necessary if new disease types or environmental conditions emerge. Comprehensive reports summarizing evaluation metrics and test conditions are key for reproducibility and peer comparison.

In conclusion, thorough and multi-dimensional evaluation ensures that machine learning models for plant disease detection are not only accurate but also reliable, efficient, and ready for deployment in diverse agricultural settings.

3. Literature Survey

A substantial body of literature has emerged over the past decade exploring machine learning techniques for plant disease detection. Mohanty et al. (2016) pioneered the use of deep CNNs on the PlantVillage dataset, achieving over 99% accuracy in classifying 26 diseases across 14 crop species. Ferentinos (2018) expanded this work by applying deeper CNN architectures and comparing results across multiple datasets. Sladojevic et al. (2016) developed a CNN model that successfully detected multiple plant diseases with high precision.

Other researchers have explored transfer learning. Too et al. (2019) utilized pre-trained models like AlexNet and ResNet to improve classification accuracy, especially on smaller datasets. Amara et al. (2017) focused on banana leaf diseases using LeNet, demonstrating the model's capability in constrained settings. Traditional ML approaches also remain relevant. Barbedo (2013) analyzed the effectiveness of handcrafted features in classical classifiers such as SVMs and Random Forests.

Multispectral and hyperspectral imaging have been studied by Behmann et al. (2015), who showed their utility in detecting early-stage symptoms. Zhang et al. (2019) implemented image segmentation and feature extraction to distinguish diseases in apples. Mohan and Nair (2020) proposed a hybrid model combining CNN features with SVM classification for tomato leaf diseases. Real-time applications have been developed, including mobile apps and drone-based detection systems.

Recent literature emphasizes robustness and interpretability. Kamilaris and Prenafeta-Boldú (2018) reviewed over 40 papers and highlighted the need for standardized benchmarks. Research by Singh et al. (2020) incorporated attention mechanisms to improve localization of disease symptoms. Emerging studies explore few-shot and zero-shot learning for rare diseases. Despite progress, challenges remain in data quality, generalization, and model interpretability.

4. Future Directions

The future of plant disease detection using machine learning lies in creating more adaptable, explainable, and scalable systems. One direction is the integration of multi-modal data, combining RGB images with hyperspectral, thermal, and contextual data. Few-shot and zero-shot learning approaches aim to address the challenge of data scarcity by enabling models to recognize unseen diseases. Federated learning allows distributed training without centralizing data, addressing privacy and collaboration challenges.

Explainable AI is critical for gaining trust among farmers and agronomists. Models should not only predict diseases but also provide visual explanations. Real-time disease detection using edge computing and Internet of Things (IoT) devices will facilitate in-field diagnostics. Integration with weather data and geographic information systems (GIS) can enhance predictive modeling of outbreaks.

Synthetic data generation using GANs offers a promise for under-represented classes. Robustness against environmental noise and image quality variations must be enhanced. Open-access, annotated, and standardized datasets will promote reproducibility and benchmarking. Interdisciplinary collaboration is needed to tailor ML models to specific agricultural ecosystems.Future systems may incorporate disease progression modeling to support proactive interventions. Personalized recommendation systems for treatment plans could be developed. Collaborative platforms that allow farmers to contribute data and receive diagnoses are likely to grow. Sustainable AI models with lower computational footprints are essential for developing regions.

Finally, government and institutional support will be crucial in implementing these technologies at scale. Addressing these future directions will enable the widespread adoption of intelligent disease detection systems and contribute significantly to sustainable agriculture.

5. Conclusion

Machine learning, especially deep learning, has revolutionized plant disease detection by offering automated, accurate, and scalable solutions. This paper reviewed traditional ML techniques, deep learning architectures, and hybrid models used for this task. It also highlighted the strengths and limitations of these approaches through comparative analysis. While impressive progress has been made, further work is needed to address data scarcity, generalization, and deployment challenges. Collaborative efforts between AI researchers, agronomists, and policymakers are essential to translate these technologies into practical tools for farmers worldwide.

References

1. S. Sladojevic et al., “Deep neural networks based recognition of plant diseases by leaf image classification,” Computational Intelligence and Neuroscience, 2016.
2. S. Mohanty et al., “Using deep learning for image-based plant disease detection,” Frontiers in Plant Science, vol. 7, p. 1419, 2016.
3. K. Ferentinos, “Deep learning models for plant disease detection and diagnosis,” Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018.
4. K. Zhang et al., “Attention-based CNN for leaf disease classification,” IEEE Access, vol. 7, pp. 42817–42826, 2019.
5. A. Amara et al., “A deep learning-based approach for banana leaf diseases classification,” Datenbank-Spektrum, vol. 17, no. 3, pp. 245–254, 2017.
6. A. Fuentes et al., “A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition,” Sensors, vol. 17, no. 9, pp. 2022, 2017.
7. E. Too et al., “A comparative study of fine-tuning deep learning models for plant disease identification,” Computers and Electronics in Agriculture, vol. 161, pp. 272–279, 2019.
8. Y. Zhao et al., “CNN-SVM: A hybrid image classification method for plant diseases,” IEEE Access, vol. 8, pp. 134395–134406, 2020.

