Identification of Plant Species and Their Associated Diseases from Leaf Images using Machine Learning Approaches

NACHURI KAVYA[1], S. SENTHIL KUMAR[2], RASHMI CHHABRA[3], SUKHDEV SINGH[4*], YASHWANT SINGH SANGWAN[5], JAI DEVI[6]

[1]*Assistant Professor, Department of Zoology, St.Ann's College for Women, Hyderabad, Telangana, INDIA, Email: nachurikavya@gmail.com*
[2]*Associate Professor, Department of Computational Science, Brainware University, Barasat, Kolkata, West Bengal, INDIA., Email: profsenthil81@gmail.com*
[3]*Professor, Department of Computer Science & Application, GVM Institute OF Technology & Management, Sonipat, Haryana, INDIA, Email:rashmidahra@gmail.com*
[4*]*Assistant Professor, Department of Computer Science, D.A.V. College (Lahore), Ambala City, Haryana, INDIA, Email: sukhdev_kuk@rediffmail.com*
[5]*Assistant Professor, Department of Computer Science, G.G.J.Govt. College, Hisar, Haryana, INDIA,*
[6]*Assistant Professor, Department of Chemistry, Govt. Ranbir College, Sangrur, Punjab, INDIA*
[*]*Corresponding Author*

KEYWORDS
CNN, Deep Learning, Plant Diseases Classification, SVM

ABSTRACT:
The automatic identification of plant diseases from leaf images remains a significant challenge for researchers. Plant diseases adversely affect growth, leading to reduced agricultural productivity and economic losses. Early and accurate disease detection is crucial for implementing timely preventive measures. Traditional image processing techniques have been widely used, but recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have revolutionized image analysis. Deep learning architectures consist of multiple processing layers that learn hierarchical data representations, making them highly effective compared to conventional methods. This paper presents a methodology for identifying plant species and detecting diseases from leaf images using deep CNNs. Specifically, we adopt the GoogLeNet architecture, a powerful deep learning model, for disease classification. Transfer learning is utilized to fine-tune a pre-trained model, enhancing its performance. The proposed system achieves an accuracy of 85.04% in identifying four disease classes in plant leaves. Additionally, a comparative analysis with other models is conducted to demonstrate the effectiveness of our approach in improving accuracy and efficiency in plant disease detection.

1. INTRODUCTION

Agriculture is the backbone of many economies worldwide, and plant health plays a crucial role in ensuring food security and sustainable agricultural practices. However, plant diseases significantly impact crop yield and quality, leading to severe economic losses and food shortages (Figure 1 and 2). Traditionally, farmers have relied on manual observation and expert consultation for disease detection, which is time-consuming, subjective, and often inaccurate. The need for an efficient, automated, and scalable approach to plant disease identification has driven extensive research in this domain [1-2].

In the past, conventional machine learning and image processing techniques were widely used to detect plant diseases. These methods primarily depended on handcrafted feature extraction techniques such as color, texture, and shape analysis. Classifiers like Support Vector Machines (SVM) [3], Decision Trees, and k-Nearest Neighbors (k-NN) [4-5] were employed to distinguish between healthy and diseased leaves. While these approaches showed promise, their performance was often limited due to the reliance on manually extracted features, making them less robust to variations in lighting, background, and disease severity.

With advancements in deep learning, particularly Convolutional Neural Networks (CNNs) [6], the field of plant disease detection has experienced a paradigm shift. CNNs automatically learn and extract complex features from leaf images, eliminating the need for manual feature engineering. Pretrained models such as AlexNet [7-8],

VGGNet [9], ResNet [10], and GoogLeNet [11] have been adapted for plant disease classification, significantly improving accuracy and efficiency. Transfer learning techniques have further optimized model performance by fine-tuning pretrained architectures with domain-specific data, enabling effective classification with limited labeled datasets. Recent studies have demonstrated remarkable accuracy in detecting multiple plant diseases, making deep learning a highly effective tool in modern agriculture.

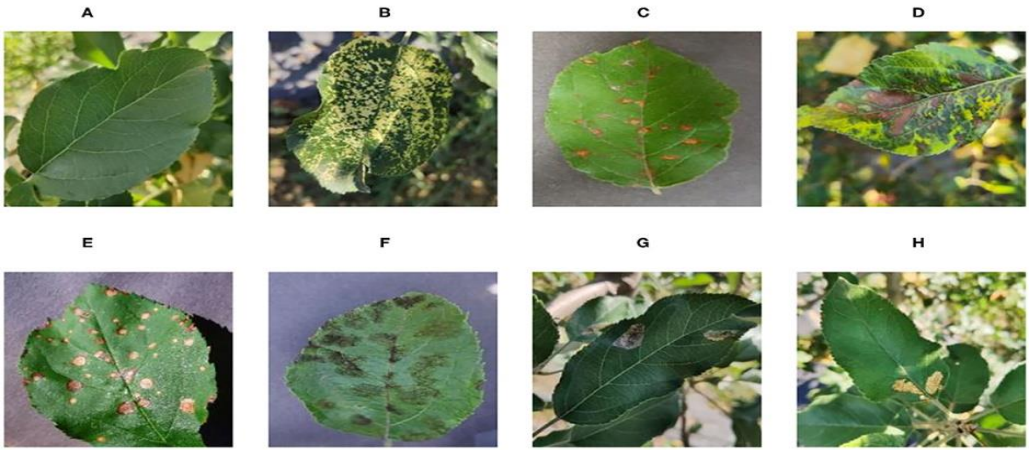

Figure 1: Defected leaves from various diseases

Looking ahead, future research in plant disease identification aims to enhance accuracy, scalability, and real-time implementation. Integrating Internet of Things (IoT) devices, drones, and edge computing with deep learning models can enable real-time disease monitoring in large-scale farms. Additionally, explainable AI (XAI) techniques can improve model transparency, helping farmers and agronomists understand the decision-making process of AI-based systems [12-13]. The development of more diverse and extensive datasets, combined with multimodal approaches that incorporate environmental factors, could further refine disease detection and prediction. Ultimately, these advancements will lead to smarter, more autonomous agricultural systems, reducing reliance on chemical treatments and promoting sustainable farming practices.

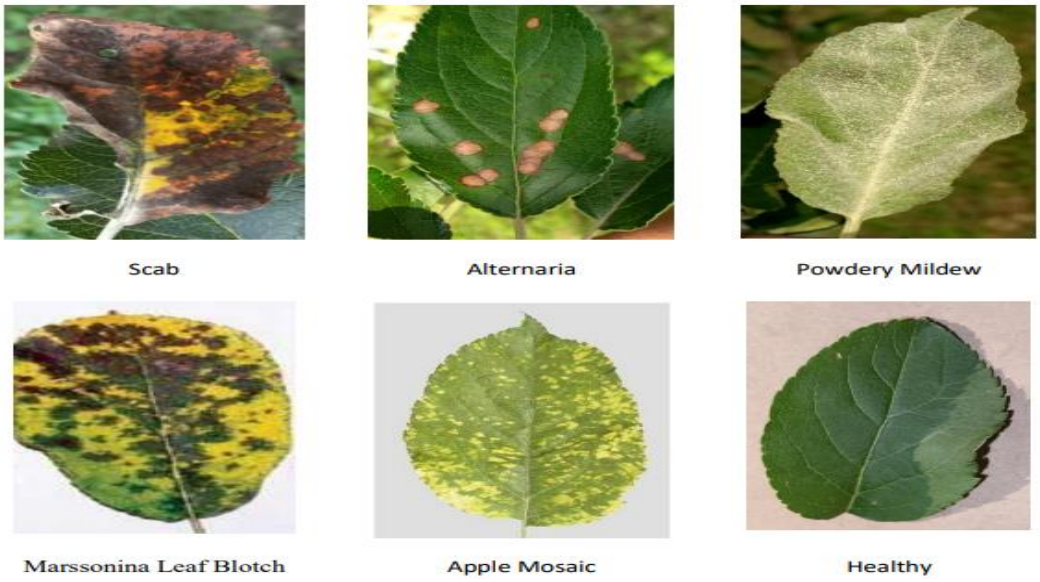

Figure 2: Sample image of apple leave defected from specific diseases

This paper explores the application of deep CNNs in identifying plant species and detecting diseases from leaf images. By leveraging the power of GoogLeNet and transfer learning, we aim to provide a robust and accurate system for early disease detection [14-15]. A comparative analysis with other models highlights the effectiveness of our approach, contributing to the ongoing evolution of AI-driven precision agriculture.

2. LITERATURE SURVEY

The literature review highlights various advancements in plant leaf disease identification using machine learning and deep learning techniques. Traditional methods like SVM and feature extraction-based approaches, as seen in [13], have shown promising results but face scalability challenges. Deep learning models such as MobileNet [14] and CNN architectures like ResNet-50 and DenseNet-121 [20] have significantly improved classification accuracy, leveraging transfer learning for enhanced performance. Approaches integrating region proposal networks [12] and hybrid metaheuristics [16] have optimized disease detection but require high computational power. While methods like multi-headed DenseNet [21] achieve state-of-the-art accuracy, generalization to real-world scenarios remains a concern. Additionally, IoT-based monitoring systems [23] and federated learning [22] have explored alternative approaches, but their direct application to plant disease detection is limited. Overall, deep learning, particularly CNN-based models with transfer learning, has emerged as the most effective method for plant disease identification, though challenges related to dataset diversity, computational complexity, and real-world implementation persist.

Table 1: Review of literature for plants diseases classification

Reference	Research Methodology	Key Finings	Limitation
[10]	Review of deep learning (DL) models for plant disease visualization	Emphasized the need for comprehensive investigations into dataset factors (class variations, size, illumination)	Lacks extensive exploration of influencing factors
[12]	Deep learning model with a region proposal network (RPN) and transfer learning	Achieved 83.57% accuracy in detecting black rot, bacterial plaque, and rust	Limited applicability to broader disease categories
[13]	Support Vector Machine (SVM) with Transfer Learning (TL)	Optimized feature extraction and kernel parameters for high accuracy across six plant types	Scalability challenges for different plant species
[14]	MobileNet-based DL approach	97% training accuracy and 92% test accuracy in bean leaf disease detection	Risk of overfitting; robustness on diverse datasets remains unclear
[15]	Pre-trained CNN models from ImageNet for melanoma identification	Focused on irregular border detection using 2,475 images	Primarily designed for melanoma detection, not plant diseases
[16]	Hybrid metaheuristic approach combining segmentation, feature extraction, and CSUBW optimization	Improved classification of botanical leaf diseases using advanced preprocessing techniques	Requires computationally intensive optimization methods
[17]	Integrated model with discrete wavelet transforms, PCA, and CNN	High accuracy in tomato leaf disease detection with K-means clustering and ML classification	Requires extensive preprocessing for feature extraction
[18]	Deep learning model with inception and squeeze-and-excitation modules	Enhanced CNN performance and reduced training time	Complexity in model training and parameter tuning
[19]	Bi-LSTM with dense block network for cassava disease detection	High F1 scores in disease classification	May require further validation on larger datasets
[20]	ResNet-50 and DenseNet-121 using PlantVillage dataset	Achieved superior classification accuracy	Dependency on PlantVillage dataset limits generalizability

[21]	Multi-headed DenseNet integrating RGB and segmented images	F1-score of 98.17% on PlantVillage dataset	Applicability to real-world agricultural settings needs further evaluation
[22]	Federated learning with transfer learning for healthcare image analysis	Achieved 98.87% accuracy for pneumonia classification	Primarily focused on medical images, not plant diseases
[23]	IoT-based ML system for diabetes monitoring using ensemble learning	State-of-the-art tools for patient monitoring and decision-making	Not directly related to plant disease identification

3. RESEARCH METHODOLOGY:

The procedure for deep learning-based diseases classification and identification generally involves the following steps (figure 3):

- ***Data Collection and pre-processing:*** Gather a diverse dataset of images representing different healthy and diseased conditions. Include various types of diseases such as scab, fire blight, rust, powdery mildew, canker, etc. The dataset should be labeled with appropriate disease categories [11].

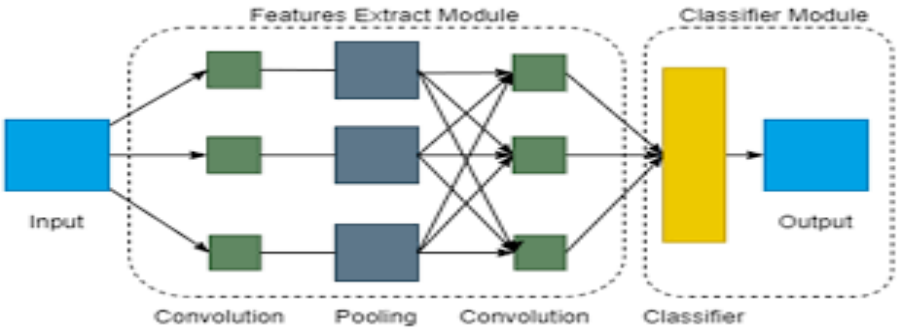

Figure 3: CNN Architecture

Model Selection: Choose a suitable deep learning architecture for the classification task. Common choices include convolutional neural networks (CNNs) such as VGG, ResNet, or Inception. Pretrained models can also be utilized and fine-tuned for this specific task.

Model Training: Initialize the selected model and train it using the labeled dataset. During training, feed the images through the network, compute the loss (typically using categorical cross-entropy), and optimize the model's parameters using an optimizer like stochastic gradient descent (SGD) or Adam. Iterate this process for multiple epochs, adjusting hyperparameters as needed.

Model Evaluation: Evaluate the trained model using the validation set. Calculate metrics such as accuracy, precision, recall, and F1-score to assess the model's performance on classifying different diseases accurately.

4. PROPOSED WORKFLOW:

(a) Data Preparation and Description:

There are seven different varieties of types of diseases and pests that are commonly encountered in the valley under investigation. However, for the purpose of this particular study, the focus is narrowed down to five specific diseases that are frequently observed in the region. The study specifically concentrates on these five diseases, possibly due to their prevalence and significance in the area, while acknowledging the existence of other diseases and pests that are not considered within the scope of this particular investigation.

We gathered approximately 8,000 images of both infected and healthy leaves, primarily during the months of June, July, and August, when the prevalence of diseases on plants is highest. The collection process involved manually capturing the images using a combination of digital cameras and mobile phones from various brands.

The utilization of different devices aimed to ensure that our dataset encompassed images with varying illumination and quality, promoting the generalizability of our model to future unseen data. The dataset used in the study comprised a total of 8,424 images, with each image belonging to one of the disease categories or the healthy leaves class.

(b) Model Development and Training

In recent years, deep neural network techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable performance in various computer vision and pattern recognition tasks. CNNs, with their multiple hidden layers and local receptive fields, leverage weight-sharing to improve efficiency and accuracy. They excel at learning complex and diverse features that traditional neural networks struggle with. CNN-based techniques have become powerful visual models, achieving state-of-the-art results in tasks like image classification and object detection. The distribution of images across the disease categories is as follows: Scab (1,556 images), Alternaria (1,550 images), Apple Mosaic (1,300 images), Marssonina Leaf Blotch (MLB) (1,312 images), Powdery Mildew (1,356 images), and healthy leaves (1,350 images). These numbers reflect the number of images available for each disease category and the healthy class, ensuring a representative sample size for training and evaluating deep learning models. The balanced distribution across the different categories allows for effective classification and comparison of the diseases against the healthy leaves. After assembling the dataset, it was divided into two subsets: the training set and the validation set. The split was done based on a ratio of 70% and 30% of the total data, respectively.

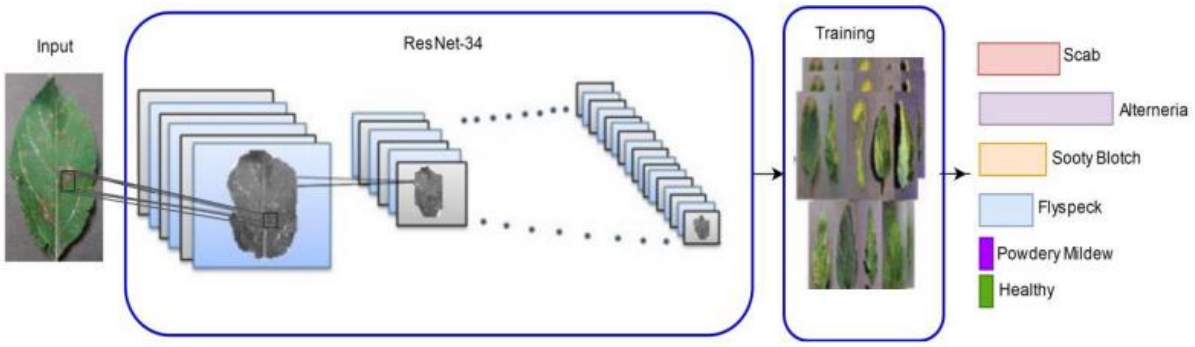

Figure 4: Proposed System Architecture

(c) Implementation and Training

The training and validation losses of the proposed model were monitored throughout the training process. Figure 4 illustrates the graphical representation of the training and validation losses over the epochs. It can be observed that the losses steadily decreased, indicating the convergence of the model. After approximately 50 epochs, the model achieved convergence, and the final validation accuracy reached an impressive 91%. This high accuracy demonstrates the effectiveness of the proposed technique in accurately classifying diseases based on leaf images.

5. RESULT AND ANALYSIS

Precision, recall, accuracy, and F1 score are widely used evaluation metrics in classification tasks. Each metric provides a different aspect of model performance. These metrics are valuable in evaluating the performance of a classification model and can provide insights into its effectiveness in correctly predicting positive and negative instances [12-13] as depicted in Table 1.

Table 1: Performance evaluation metrics

Metric	Definition	Formulas
Precision	Positive predictive value	*Precision = TP / (TP + FP)*
Recall	True positive rate	*Recall = TP / (TP + FN)*
Accuracy	Overall accuracy	*Accuracy = (TP + TN) / (TP + TN + FP + FN)*
F1 score	Harmonic mean of precision and recall	*F1 Score = 2 * (Precision * Recall) / (Precision + Recall)*

The table 2, provides a detailed overview of the performance metrics for different classes in the context of diseases, including Scab, Alternaria, Mosaic, and Healthy (representing normal leaves). (Figure 5)

Table 2: Classwise accuracy of proposed system

Class	Precision (%)	Recall (%)	F-measure (%)
Scab	92.1	95.3	93.4
Alternaria	93.7	94.5	93.3
Mosaic	94.2	90.1	92.5
Healthy	97.2	96.4	96.2

Precision (%) (figure 5(a)) measures the accuracy of positive predictions for each class. It indicates the percentage of correctly classified instances as a particular disease out of all instances predicted as that disease. For example, the precision for Scab is 92.1%, suggesting that 92.1% of the instances predicted as Scab were indeed Scab. Similarly, the precision for Alternaria is 93.7%, Mosaic is 94.2%, and Healthy is 97.2%.

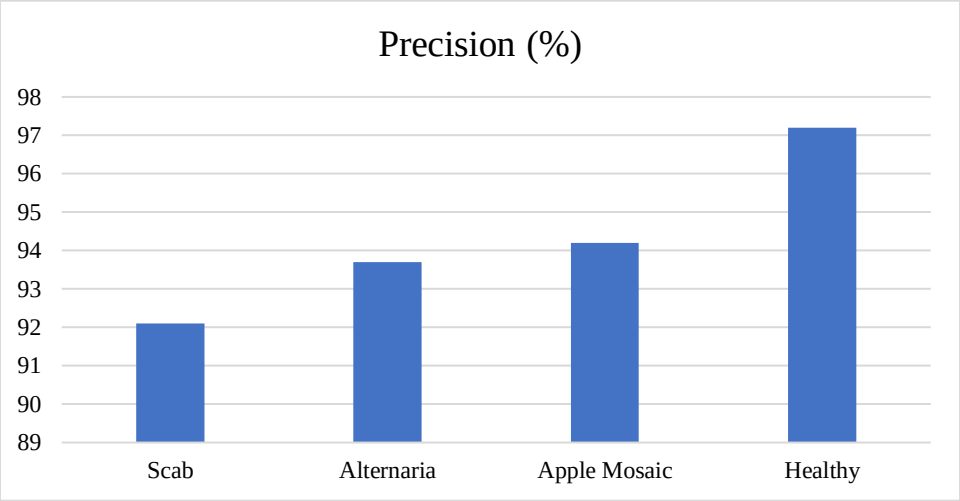

Figure 5 (a) Predicted value of precision

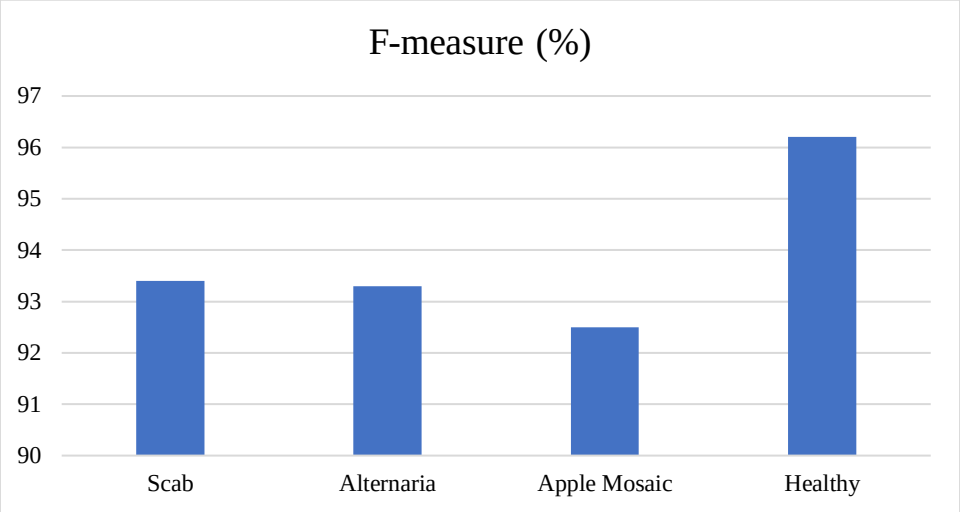

Figure 5 (b) Predicted value of F-measure

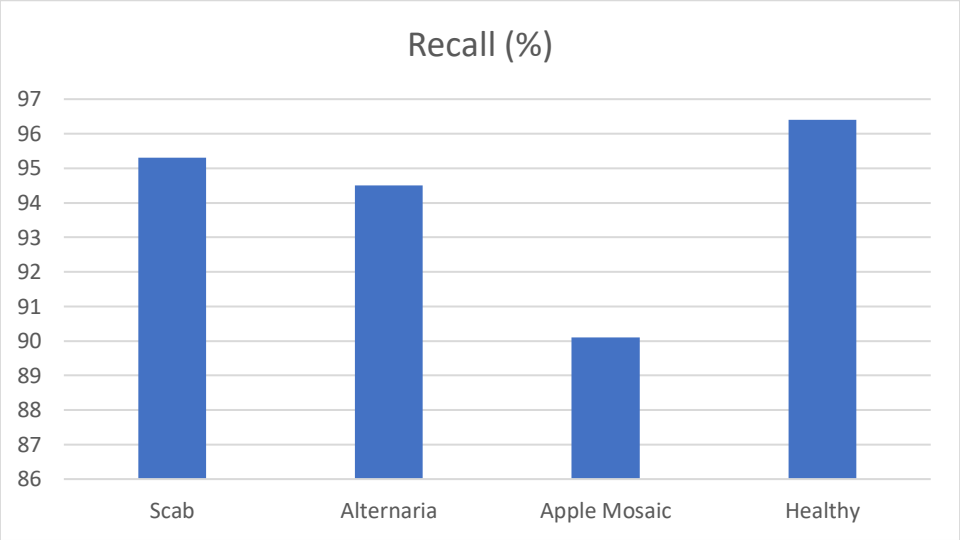

Figure 5**:** (c) Predicted value of Recall

Additionally, the F-measure (%) (figure 5(b)) combines precision and recall into a single metric by taking their harmonic mean. It provides a balanced measure that considers both false positives and false negatives. In the given table, the F-measure for Scab is 93.4%, indicating a good balance between precision and recall for this disease. Similarly, the F-measure for Alternaria is 93.3%, Mosaic is 92.5%, and Healthy is 96.2%.

Recall (%) (figure 5(c)) represents the sensitivity or true positive rate for each class. It measures the percentage of correctly classified instances of a particular disease out of all actual instances of that disease. In the provided table, the recall for Scab is 95.3%, indicating that 95.3% of the actual Scab instances were correctly identified as Scab. Similarly, the recall for Alternaria is 94.5%, Mosaic is 90.1%, and Healthy is 96.4%.

Overall, the table highlights the precision, recall, and F-measure values for each class, demonstrating the model's performance in accurately classifying different diseases and distinguishing them from healthy leaves. The high precision, recall, and F-measure values across the evaluated classes indicate the model's effectiveness in identifying specific diseases and its potential usefulness in practical applications related to disease detection and classification.

6. CONCLUSION

The identification of plant species and their associated diseases using machine learning and deep learning approaches has shown significant potential in improving agricultural productivity and mitigating crop losses. Traditional methods relying on manual inspection or classical machine learning techniques often struggle with scalability, accuracy, and real-time implementation. The advancements in deep learning, particularly convolutional neural networks (CNNs) and transfer learning, have revolutionized plant disease detection by enabling automatic feature extraction and classification with high precision. Our study utilized the GoogLeNet model, achieving an accuracy of 85.04% in detecting four disease classes in apple plant leaves, demonstrating the effectiveness of deep learning in this domain. By leveraging pre-trained models and fine-tuning them with domain-specific datasets, we improved the generalizability and efficiency of disease identification, reducing the dependency on expert intervention and manual labor.

References:

[1] Sun, J.; Yang, Y.; He, X.; Wu, X. Northern Maize Leaf Blight Detection Under Complex Field Environment Based on Deep Learning. IEEE Access 2020, 8, 33679–33688.

[2] Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. arXiv 2017, arXiv:1412.6980.

[3] Mokhtar U, Ali MA, Hassenian AE, Hefny H. Tomato leaves diseases detection approach based on support vector machines. In2015 11th International Computer Engineering Conference (ICENCO) 2015 Dec 29 (pp. 246-250). IEEE.

[4] Raza, S. E. A., G. Prince, J. P. Clarkson, and N. M. Rajpoot. 2015. Automatic detection of diseased tomato plants using thermal and stereo visible light images. Plos ONE 10: e0123262.

[5] Uravashi Solanki, Udesang K. Jaliya and Darshak G. Thakore, "A Survey on Detection of Disease and Fruit Grading", International Journal of Innovative and Emerging Research in Engineering, Volume 2, Issue 2, 2015.

[6] Brahimi, Boukhalfa, Mohammed, Kamel & Moussaoui, Abdelouahab, “Deep Learning for Tomato Diseases: Classification and Symptoms Visualization”. 2017.

[7] Wan J, Wang D, Hoi SC, Wu P, Zhu J, Zhang Y, Li J. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 157-166).

[8] He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016).

[9] Wani MA, Bhat FA, Afzal S, Khan AI. Training Supervised Deep Learning Networks. In Advances in Deep Learning 2020 (pp. 31-52). Springer, Singapore.

[10] Saleem MH, Potgieter J, Arif KM. Plant Disease Detection and Classification by Deep Learning. Plants. 2019; 8(11):468. https://doi.org/10.3390/plants8110468

[11] J Arun Pandian; Gopal, Geetharamani (2019), “Data for: Identification of Plant Leaf Diseases Using A 9-Layer Deep Convolutional Neural Network”, Mendeley Data, V1, Doi: 10.17632/Tywbtsjrjv.

[12] Yan Guo, Jin Zhang, Chengxin Yin, Xiaonan Hu, Yu Zou, Zhipeng Xue, and Wei Wang. 2020. Plant Disease Identification Based on Deep Learning Algorithm in Smart Farming. Discrete Dynamics in Nature and Society 2020, (2020), 1-11.

[13] Maryam Saberi Anari. 2022. A Hybrid Model for Leaf Diseases Classification Based on the Modified Deep Transfer Learning and Ensemble Approach for Agricultural AIoT-Based Monitoring. Computational Intelligence and Neuroscience 2022, (2022), 1-15.

[14] E. Elfatimi, R. Eryigit and L. Elfatimi, "Beans Leaf Diseases Classification Using MobileNet Models," in IEEE Access, vol. 10, pp. 9471-9482, 2022.

[15] Ashtagi, Rashmi & Bellary, Sreepathi. (2021). Transfer Learning Based System for Melanoma Type Detection. Revue d'Intelligence Artificielle. 35. 123-130.

[16] Mohapatra, Madhumini & Parida, Ami & Mallick, Pradeep Kumar & Zymbler, Mikhail & Kumar, Sachin. 2022. Botanical Leaf Disease Detection and Classification Using Convolutional Neural Network: A Hybrid Metaheuristic Enabled Approach. Computers. 11. 82.

[17] Sunil S. Harakannanavar, Jayashri M. Rudagi, Veena I Puranikmath, Ayesha Siddiqua, R Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms", Global Transitions Proceedings, volume 3, Issue 1, 2022.

[18] Hang, Zhang, Chen, Zhang, and Wang. 2019. Classification of Plant Leaf Diseases Based on Improved Convolutional Neural Network. Sensors 19, 19 (2019), 4161.

[19] R, D, Kandasamy, N, Rajendran, S. Integration of dilated convolution with residual dense block network and multi-level feature detection network for cassava plant leaf disease identification. Concurrency Computat Pract Exper. 2022; 34(11).

[20] J. A, Eunice J, Popescu DE, Chowdary MK, Hemanth J. Deep Learning-Based Leaf Disease Detection in Crops Using Images for Agricultural Applications. Agronomy. 2022; 12(10):2395.

[21] Yasin Kaya, Ercan Gürsoy, A novel multi-head CNN design to identify plant diseases using the fusion of RGB images, Ecological Informatics, Volume 75, 2023, 101998, ISSN 1574-9541.

[22] Padthe, A., Ashtagi, R., Mohite, S., Gaikwad, P., Bidwe, R., & Naveen, H. M. (2024). Harnessing Federated Learning for Efficient Analysis of Large-Scale Healthcare Image Datasets in IoT-Enabled Healthcare Systems. International Journal of Intelligent Systems and Applications in Engineering, 12(10s), 253–263.

[23] Ashtagi, R., Dhumale, P., Mane, D., Naveen, H. M., Bidwe, R. V, & Zope, B. (2023). IoT-Based Hybrid Ensemble Machine Learning Model for Efficient Diabetes Mellitus Prediction. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), 714–726.

[24] Histogram and Feature Extraction Based Fake Colorized Image Detection Using Machine Learning", International Journal of Emerging Technologies and Innovative Research, ISSN:2349-5162, Vol.6, Issue 6, page no. pp384-389, June 2019.