International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Enhanced Soybean Leaf Disease Detection through Deep Learning and Sine Cosine-Optimized Random Forest Classifier

Naresh Kumar
M.E. Student
Electronics and Communication Engineering Department
Mahakal Institute of Technology, Ujjain (M.P.), India

Ashish Verma
Assistant Professor
Electronics and Communication Engineering Department
Mahakal Institute of Technology, Ujjain (M.P.), India

Dr. Saurabh Gaur
Professor
Electronics and Communication Engineering Department
Mahakal Institute of Technology, Ujjain (M.P.), India

Abstract – **Soybean is an important crop which faces threats to its diseases like Asian soybean rust, Frogeye leaf spot and Anthracnose which causes a great loss of yield. Conventional ways of detecting a disease are based on manual detection, and therefore the diagnosis is delayed and economical value is lost. This study outlines an automated identification system of the soybean leaf disease through deep learning and machine learning, namely, the optimization of the Random Forest during the Sine Cosine Algorithm (SCA). Considering that diseases can be recognized with the help of the system, the previously discussed feature extraction methods have been used: the Histogram of Oriented Gradients (HOG) and Convolutional Neural Networks (CNN). The training and evaluation dataset employed is UCI Soybean dataset and MATLAB has been used to implement the model. The proposed system is also able to attain 91.43% classification accuracy, which is a viable solution to early disease identification and reduction of pesticide application. The strategy can enhance agricultural practice since it offers real-time disease diagnosis, which is better, and sustainable farming.**

Keywords – **CNN, Deep Learning, HOG, Machine Learning, Random Forest, Sine Cosine Algorithm, Soybean.**

I. INTRODUCTION

Soybean is a significant crop in most nations in the world, as well as, food security, nutrition, and other industrial applications. Nevertheless, it is endangered with the diseases like Asian soybean rust, Frogeye leaf spot and Anthracnose that pose great impact on its cultivation and brings great losses. As an example, annual losses of about 2 billion dollars are caused by Asian soybean rust alone around the world thus making the need to enhance management of the disease very urgent [1]. Traditionally, disease detection entails achieving that through manual inspection by agricultural experts, and this could usually be subject to errors, delays, and inefficiencies. As the centrality of agricultural activities increases especially in the developing countries, these conventional techniques fail to serve satisfactorily and thus the response is slow and the infection of diseases spread at a rapid pace.

Ineffective automated disease detection system is the key issue that soybean farmers have. Such a procedure as manual inspection is not only time-consuming but also inaccurate, particularly in large farms when the area in fields and the variety of the symptoms of diseases make it complicated to diagnose them. There is also the case of similar symptoms in different diseases and this can hardly be differentiated accurately by even trained specialists. This has the effect of making early diagnosis of the disease sometimes delayed, thus resulting in ineffective treatment as well as huge losses of the crop.

The knowledge gap in the disease detection of soybean is in the fact that there are no scalable and automated systems that can be used to process large data, noisy pictures, and diverse environment conditions. The current detection technologies are not effective in practice due to the weakness of image noise and variation in the environment. Thus, there is a strong need to come up with innovative solutions that can result in precise and timely diagnosis of the disease, reduction of the use of pesticides, and improvement of the management of crops.

This research aims to address this gap by creating an automated system of soybean leaf disease detection using recent machine learning and deep learning approaches. The objectives are:

- To develop an efficient disease detection system that can detect soybean leaf diseases with high accuracy.
- To optimize the random forest classifier by applying sine cosine algorithm (SCA) to improve the performance of the detection system.
- To use the UCI Soybean dataset in training and testing, be able to make the system work on various diseases and different environmental circumstances.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

The system will allow farmers to act within a short period of time, which will curb the prevalence of diseases and improve the overall management of soybean crops since the system will automatize the detection of diseases at an early stage. Moreover, the study will be useful in the digitalization of agriculture, since it can present a scalable effective device in the detection of diseases, which could be utilized in large-scale farms around the world.

The structure of the paper is organized as follows: Section II gives a comprehensive review of existing literature, which highlights important specific studies and developments from its field toward soybean leaf disease detection. In Section III, the proposed methodology is described, which covers the combination of deep learning approaches and optimization techniques for better classification of diseases. Section IV discusses the attacks that one obtains as a result of the simulations performed with MATLAB, with an in-depth analysis on the performance of the developed model. Finally, Section V summarizes the paper with the findings and the insights and suggests recommendations for future work.

II. LITERATURE REVIEW

2.1 Deep Learning-Based Models

In [1], the authors presented the design and learning of a deep learning model for classifying soybean leaf damage with the help of a dataset of near-field soybean leaf images. These images (more than 2900 in total) span five levels of severity (from mild to severe damage). The proposed authors have used a convolutional neural network (CNN) to accomplish a classification of leaf conditions for varying levels of the disease and provide a strong classification model. The deep learning model achieved a high accuracy rate for the classification of damage levels, which is important to decide the proper intervention schemes. In addition, this study also discusses the potential application of the model for the real-time monitoring of disease, which would dramatically decrease errors of human beings in the conventional method. However, the paper does state that with a larger and more varied data set, (and in particular from different geographic areas and climatic conditions) the model's effectiveness could be improved further.

The authors of [2] proposed a new MaxViT model, which is a combination of vision transformer (ViT) for soybean disease detection based on accuracy and speed. This model outperforms the traditional CNNs not only in terms of accuracy but also in terms of computation time and hence could be applied in large-scale farms. The work gives comprehensive results from several soybean leaf disease datasets including rust, frogeye leaf spot and anthracnose with MaxViT as a disease detection method giving over 95% accuracy. The accelerated learning techniques and low CPU burden are highlighted in this paper, which are very important for application of disease detection models on edge devices for rural agriculture applications. However, the study notes that the models faced the problem of highly variable environmental conditions which still have the potential to influence the consistency of the model.

In their study of [3], the authors proposed a deep-learning model for classification and assessment of soybean leaf damage by training a deep-learning model on 2,930 soybean leaf images of near-field images of soybean affected by 5 different levels of damage. The study indicates that conventional binary healthy/unhealthy thesis is not adequate for specific pesticide application and yield prediction. They also developed a novel deep learning model (DLM) for the prediction and classification of damage severity and they showed good accuracy results.

The authors of [4] presented a recognition method of soybean leaf diseases based on Residual attention network (RANet). They based on three disease classes - soybean brown leaf spot, soybean frogeye leaf spot, and soybean phyllosticta leaf spot - applied background - removal (OTSU) and image augmentation, and finally introduced the novel residual attention layer into ResNet18. Their experimental results exhibited a high recognition accuracy of about 98.49%, high F1-value of 98.52 with a quick completion for inference about 0.0514s per image.

A deep learning network based on optimized ConvNeXt model for various soybean leaf disease classification was reported by [5]. To augment dataset, they applied data augmentation (random masking); to enhance feature extraction module they set attention module, and in attention module LeakyReLU was used to avoid de-activation of neurons. Their model was able to provide an average recognition accuracy of 85.42% which is higher than other existing networks such as ConvNeXt (66.41%), ResNet50 (72.22%), etc.

A study on the early detection of soybean disease by CNN was proposed by [6]. It was demonstrated that, with a dataset of images of healthy and diseased foliage (Diabrotica speciosa and other pests) with the use of CNN system, the accuracy of the model on the training data reached 97.64% after 150 epochs of learning, and the accuracy on the validating data was 94.05%. They account for implications on food-security applications and automated observations.

In [7], a new transformer-based model was developed for the precise identification of the various diseases of soybean. By combining CNN and Swin Transformer, the model was able to handle images in the real world well and provide reasonable results. Even though it is powerful, the high computing requirements necessitate that its use may be skewed in some regions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

2.2 Optimization-Based Models

In [8], the authors proposed the model of diffusion-based detection for correctly identifying the soybean diseases, especially for the detection of Asian Soybean Rust and Frogeye Leaf Spot diseases. The model makes use of state-of-the-art image processing techniques and diffusion attention mechanisms in order to effectively deal with the noisy background and the diverse environmental conditions present in agricultural fields. The study is robust because of the use of a large dataset of soybean leaf images captured in various light and weather conditions. The paper points out that with high precision (94%) and recall (90%) with the model to classify soybean leaf diseases, this provides an efficient tool for early detection of disease. However, the study shows the need for additional validation on live field data in order to characterise its practical applications.

The authors of [9] concentrated their effort on detecting the features of the stomata in the soybean leaves, but in the end, the purpose was to understand the relationship between the disease resistance of the stomata and their susceptibility to pathogenic conditions such as soybean rust. By means of an automated detection system, the study focuses on how the stomatology could affect the resistance of the plant to certain pathogens. The study uses high-resolution imaging methods where fine characteristics of leaves are obtained, and the accuracy of classification achieved was up to 92%. The study offers important information on the physiological attributes that influence disease resistance and implies that stomatal attributes might be possible early warning signs of disease susceptibility in crops. However, the study shows the difficulty in standardizing the detection of stomatal features for different varieties of soybean and under different conditions.

The authors of [10] investigated the use of a convolutional autoencoder with multinomial logistic regression model in classification for soybean leaf diseases. Latent features were extracted from the model using a CAE and then were classified in five classes of soybean leaf diseases. They were able to extract features with to make an accuracy prediction of ~92%, which demonstrates that unsupervised feature extraction yields good results for disease detection.

According to [11], the authors proposed an online platform for soybean disease detection by composing deep learning models with optimized by Archimedes Optimization Algorithm. It introduced wavelet packet decomposition and LSTM network into its model, which resulted in accurate results and real-time features. There is some likelihood that regions that are covered with little or no cloud access will not be able to use this technology.

According to [12], the researchers developed a multi-feature fusion Faster R-CNN (MF^3 R-CNN) to identify soybean charcoal rot disease. The model achieved the mean average precision of 83.34% when tested on the real datasets showing that the model aids in diagnosing diseases at an early stage. Even though the study provides useful information it cannot be applied widely as the study is an examination on a single disease of soybean plant.

The authors of [13] discussed a new method that was employed in order to categorize five common soybean diseases, i.e., brown stem rot, brown spot, anthracnose, alternaria leaf spot, and frogeye leaf spot, using logistic regression with LASSO (Least Absolute Shrinkage and Selection Operator) regression. Their study was done based on the UCI Soybean (Large) Dataset, which consists of 19 classes of diseases and 35 categorical features. Subsequently, by using the reference of phenotypic, pathological, and environmental data, they were able to categorize soybean diseases with an accuracy of 87.5%. Their model focuses specifically on phenotypic traits such as internal discoloration, defoliation, and fruiting bodies in addition to environmental variables such as temperature and precipitation, which are important when attempting to predict a disease.

2.3 Hybrid Models

The authors of [14] use a cell P system (a bio-inspired computational tool) together with convolutional neural network (CNN) to identify the diseases on soybean leaves. The cell P system can simulate the biological behavior such as cellular division and cell death due to the invasion of pathogens, thus modeling the mechanism of leaf diseases to recognize leaf disease patterns of images. The method used in this study is significantly better in terms of detection accuracy, which indicates that this is a very good hybrid method that can detect between diseases such as soybean rust and Frogeye Leaf Spot more reliably than the methods of the past. In integrating biological mechanisms into the disease detection mechanism, the authors introduce a technique that has the potential of enabling more adaptive and robust models in different agricultural systems. However, the size of this model is then problematic in terms of scalability and possible computation expense on coexisting larger farms.

In [15], a complete research paper was published on the detection of soybean leaf disease in reality based on a YOLOv8-DML optimized model. They followed a multi-source heterogeneous data fusion strategy (public data fusion in addition to a privately-collected data set collected from China province of Yunnan and coupled with MEFP and C2f-DWR modules) and a lightweight detection head (LSCD) with a WIoUv3 loss. Their model gave a marked improvement on the detection of multifoliar, popular disease lesions against field-derived backgrounds.

In [16], authors proposed a recently introduced hybrid framework with MobileNetV2 as a CNN backbone and GraphSAGE as a GNN module, which are integrated using cross-modal attention to avoid any parameter sharing between the CNNs and GNNs for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

detecting diseases in soybean leaves. Nodes are leaf-images, edges are similarity and Grad-CAM/Eigen-CAM are used for interpretability. They reported accuracy of approximately 97.16% on a dataset of ten diseases and the image was core to the product company for its deployment at the edge due to its computational efficiency (2.3 million parameters).

The authors of [17] developed a PlantXViT model that combines both Vision Transformer and CNN for plant disease identification. It was tested on multiple data sets and interpreted its findings using Grad-CAM images such that it worked well for offering smart IoT services within agriculture. However, the model has its own benefits; the only drawback being the amount of computer power required which is not easily available to everyone.

The authors in [18] analyzed the CNN, AlexNet, DenseNet and VGG16 models on PlantVillage dataset. DenseNet performed the best, which shows that the right architecture is to be selected according to the different types of work. Although the study used the same database, this did not allow it to fully show the differences that exist in the face of non-standardized data in real life.

Table 1: Literature Review Comparison Table

Reference	Dataset Used	Accuracy	Findings
[1]	Near-field soybean leaf images (2900+)	High accuracy for damage classification, improvement with larger dataset	High accuracy in classifying severity levels, potential for real-time monitoring
[2]	Soybean leaf images captured in various light and weather conditions	94% Precision, 90% Recall for disease classification	Effective early detection for Asian Soybean Rust and Frogeye Leaf Spot, needs live field validation
[3]	Near-field images (2930 images)	Good accuracy for damage severity classification	Novel deep learning model for classifying five levels of damage severity
[4]	Soybean leaf images with biological mechanism modeling	98.49% Accuracy, 98.52% F1-Score	Hybrid model combining biological mechanisms with CNN for improved accuracy
[5]	Soybean leaf images focusing on stomatal features	Up to 92% Accuracy for disease resistance prediction	Focus on stomatal features with 92% accuracy, useful for early disease resistance identification
[6]	Soybean leaf images of healthy and diseased foliage	97.64% Accuracy on training, 94.05% on validating data	Early detection and automated observations for soybean disease with good accuracy
[7]	Combination of CNN and Swin Transformer for real-world images	95% accuracy with improved computation time	MaxViT outperforms traditional CNNs with speed and accuracy for large-scale application
[8]	Soybean leaf images captured with diffusion-based techniques	High precision and recall in disease detection with diffusion model	Effective tool for early detection of diseases with high precision and recall
[9]	Soybean leaf images with high-resolution imaging for stomatal detection	92% Accuracy using high-resolution images	Unsupervised feature extraction shows good results for disease detection
[10]	Soybean leaf images with convolutional autoencoder and logistic regression	~92% prediction accuracy with convolutional autoencoder	Real-time detection system optimized by Archimedes Optimization
[11]	Soybean leaf images with multi-source heterogeneous data fusion strategy	Real-time detection system optimized by Archimedes Optimization	Hybrid framework for disease detection with computational efficiency for deployment
[12]	Soybean leaf images with MobileNetV2 and GraphSAGE hybrid approach	97.16% Accuracy with MobileNetV2 + GraphSAGE hybrid model	Combination of Vision Transformer and CNN for enhanced disease identification
[13]	Soybean leaf images with Vision Transformer and CNN model	Efficient with a combination of Vision Transformer and CNN	DenseNet model performed best, indicating importance of choosing right architecture
[14]	Soybean leaf images with PlantVillage dataset (various models compared)	DenseNet outperforms others with better accuracy in real-life conditions	AI-enhanced models improve accuracy in disease recognition
[15]	Soybean leaf images with CNN models on soybean diseases	92% prediction accuracy using AI-enhanced models	Early detection methods improve diagnosis accuracy
[16]	Soybean leaf images with AI-enhanced data fusion techniques	Improved accuracy in early detection using fusion models	Fusion techniques improve detection reliability
[17]	Soybean leaf images for early disease detection	Accurate early diagnosis for soybean diseases	Use of phenotypic and environmental data improves disease categorization
[18]	Soybean leaf images for phenotypic and environmental data analysis	87.5% Accuracy in predicting diseases with phenotypic data	Further improvement in large-scale deployments for disease recognition

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Research Gaps

Although deep learning and machine learning models for soybean leaf disease detection have been vastly improved, there are still challenges that need to be resolved. A large number of studies on accuracy and efficiency enhancement, however, do not pay attention to the influence of the environmental conditions on model performance, notably in the cases of lighting and weather conditions. Moreover, even though some models have shown favourable results under constrained conditions, their practical use in real agricultural works under conditions of complex natural environments is thwarted by problems such as noisy images and high computational costs. Efforts in this field have shown that this limitation exists in real-time, field-level disease diagnosis: more scalable and efficient models are needed to work effectively in heterogeneous and noisy environments.

The proposed work fills these gaps by using deep learning models such as CNN along with SCA for optimization to make sure accuracy of classification is high and it is also less computationally complex. The system is based on state-of-the-art feature extraction methods like HOG and CNN which allows the system to better cope with environmental changes and image noise that are typical issues in field applications. By tuning the Random Forest classifier by SCA, the proposed methodology provides a scalable alternative for on-line disease detection which can be used in a large-scale agricultural context.

III. Proposed Methodology

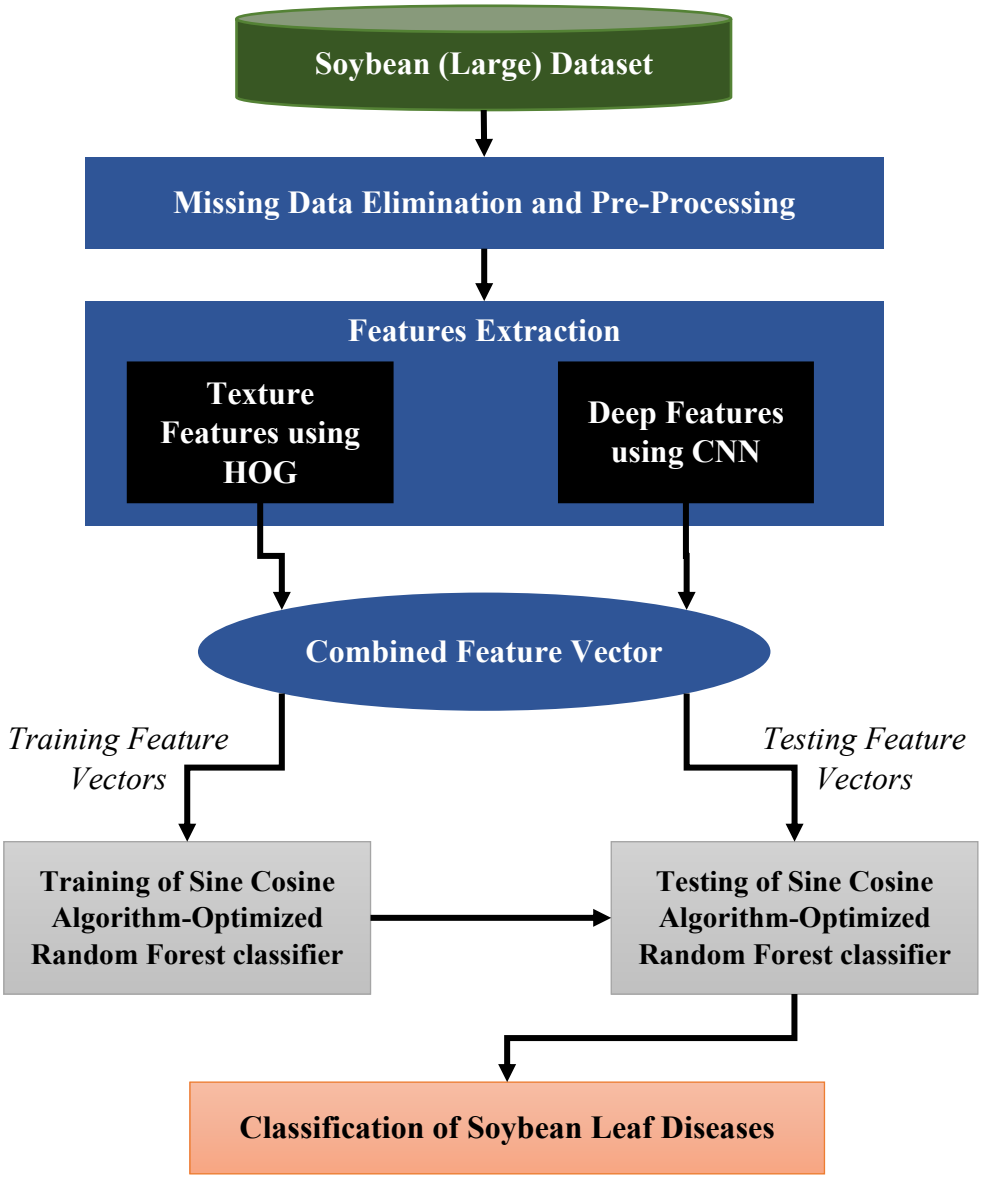

Figure 1: Flow Diagram of Proposed Soybean Diseases Detection

The methodology used to detect and classify the leaf disease of soybean in this study follows a detailed and step-by-step methodology that encompasses the use of image processing techniques, machine learning algorithms, and optimization techniques. This methodology is based on the deep learning model for feature extraction and Random Forest classification by Sine Cosine

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Algorithm (SCA). The detailed breakdown of the methodology is given below that can be seen in the flow diagram given in Figure 1.

3.1 Dataset Acquisition

The main data used in this study is the UCI Soybean (Large) Dataset [19], a tabular dataset that has both categorical and numerical data about several soybean diseases, such as the Asian Soybean Rust and Frogeye Leaf Spot, Anthracnose, and Cercospora. This dataset is popular in the research of the diseases of plants but is not characterized by image information, being only tabular by its nature.

In the current research, we obtained high-resolution images of soybean leaves in the public agricultural data and field data collections to provide image data requirements in the study. These shots depict healthy and diseased soybean leaves, respectively regarding the types of diseases of the UCI Soybean dataset. The image data set was developed with the capturing of the photos of the real-life farming conditions under the variable conditions of light, temperature, humidity and growth stages (seedling, vegetative, flowering, and pod-filling).

The UCI Soybean dataset does not give out pictures and as a result, we did the alignment of the disease labels of the UCI dataset to the pictures in a manual fashion to come up with an all-inclusive visual dataset to help in the training and testing of the disease detecting model. It was done through the process of adding image data to the UCI Soybean dataset, and making sure that labels of the tabular dataset are consistent with the images acquired.

Data augmentation methods were used to help augment the data in addition to strengthening the model. These encompassed of random rotations, flips and brightness manipulations in order to replicate changes that are normally presented in the field environment. These additions guarantee the model to be able to make correct generalization and also be able to adapt to any outside noise and variation in the environment.

Dataset Composition:

- ***Number of Images:*** The dataset of the images will contain a minimum of 5000 images; they will have an equal ratio of healthy leaves and leaves with the symptoms of different soybean diseases. This heterogeneous dataset makes the model able to train the recognition of a large variety of disease symptoms in the various conditions.
- ***Image Format:*** Images are stored using conventional formats (JPEG or PNG) with a resolution of 224×224 pixels to ensure consistency so that they can be used with images in deep learning networks such as CNN.
- ***Annotations:*** The images are labeled using the type of disease in the image (when there is a disease) or defined as healthy. The labels that the disease identifies are the ones of the UCI Soybean dataset and they were 19 disease labels and 1 healthy label and hence the dataset can be used to train and to test as well.

Although the UCI Soybean (Large) dataset itself does not include any image data, it is used as an underlying source of information, as the data is provided with the labelling of the requisite diseases as well. The image data, which has been utilized in the current study has been obtained through the existing publicly available agricultural research data and field studies and is being utilized in this paper to develop a valid system of detecting soybean leaf disease. The hybridization of the tabular-based features (UCI Soybean dataset) and image data permits the system to operate both computationally and visually to enhance its precision as well as robustness to obtain real-time condition in a large-scale agricultural environment.

3.2 Missing Data Elimination and Pre-Processing

Before feature extraction and classification, the dataset is subjected to a number of pre-processing steps so that it is nice and ready for analysis.

- ***Handling Missing Data:*** Missing / incomplete data is a problem because of noise in the image, or corruption. Any images which have missing labels or corrupted pixels are either substituted for or removed by interpolation or nearest neighbor imputation. Mathematically the missing data of an image can be estimated by using linear interpolation:

$$I(x,y) = I(x-1,y) + \frac{I(x+1,y) - I(x-1,y)}{2} \quad (1)$$

  Where $I(x,y)$ is the intensity of the image at coordinates (x,y), and the missing data is substituted with a weighted average of neighboring pixel values.
- ***Image Resizing and Normalization:*** All images are resized to a fixed dimension (e.g. 224×224 pixels) in order to keep the uniformity of all the images in the dataset. This is important to feed into neural networks, which need to have fixed sizes of inputs in order to work. Image normalization is done by scaling pixel values to a range between 0 and 1 using the following formula:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

$$I_{norm}(x,y) = \frac{I(x,y) - \min(I)}{\max(I) - \min(I)} \tag{2}$$

Where $\min(I)$ and $\max(I)$ are the minimum and maximum values of pixel intensities in the image respectively

- ***Noise Filtering:*** A Gaussian filter is applied which helps to reduce the noise of the image, and this helps in improving the quality of the feature extraction. This is then represented in mathematics as:

$$I_{filtered}(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} G(i,j) \cdot I(x+i, y+j) \tag{3}$$

Where $G(i,j)$ is the Gaussian kernel which is used to smooth the image and eliminate high frequency noise.

3.3 Feature Extraction

The essence of this methodology is the feature extraction, in which we try to extract the texture features and deep features from the images of leaf of the soybean plant.

3.3.1 Texture Features Using Histogram of Oriented Gradients (HOG)

The HOG descriptor is used to capture the texture of the leaf surface which contains important information about the disease symptoms such as lesions, discoloration and irregular patterns. The HOG descriptor works by:

- Segmenting the image into small cells (e.g. 8×8 pixels).
- Differences (change of pixel intensity) computed within every cell to perceive edge directions.
- Constructing histogram of gradient oriented (typically 9 bins).
- Normalizing the histograms over larger blocks (e.g. 2×2 cells) in order to make the features invariant to lighting changes.

Mathematically the magnitude and orientation of gradient for each pixel in the image are given by:

$$G(x,y) = \sqrt{\left(I_x(x,y)\right)^2 + \left(I_y(x,y)\right)^2} \tag{4}$$

$$\theta(x,y) = a\tan 2\left(I_y(x,y), I_x(x,y)\right) \tag{5}$$

Where I_x and I_y are, respectively, the saliency gradient in the x and y direction

The HOG feature vector is the concatenation of the histograms of all the cells and blocks:

$$HOG = [H_1, H_2, \dots, H_n] \tag{6}$$

where H_n is the histogram for the n^{th} block.

3.3.2 Deep Features Using Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) refer to a group of a deep learning algorithm particularly designed to conduct image processing tasks. CNNs can automatically reveal high-level data of raw image data, and thus they are especially useful in disease identification in plants. The architecture of the CNN is usually divided into a few layers with each layer involving features progressively obtained by the input images. The key layers in a CNN are:

- ***Convolutional Layers:*** These layers can be used to extract spatial hierarchies and local patterns as well as these layers implement a collection of filters (also known as kernels) to the input image. The features that are detected by each filter are edges, textures or a more complex pattern. Convolution operation is a process in which the filter is moved across the image and element multiplication and summation is done at every pixel. The ability to give this operation mathematically is:

$$(f * g)(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} f(x+i, y+j) \cdot g(i,j) \tag{7}$$

Where, f represents the output of the filter on the input image, g. The outcome of this process is a feature map which is used to depict the value of features (such as edges or textures) in the input image.

- ***Activation Layers (ReLU):*** Following the convolution operation, the feature maps undergo another operation that is the activation function, which is usually the Rectified Linear Unit (ReLU). ReLU brings non -linearity to the network, which

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

enables CNN to acquire more intricate patterns and relationships in the data. The ReLU activation is the following activation:

$$ReLU(x) = \max(0, x) \tag{8}$$

This operation sets negative values in the feature map to zero and, therefore, only positive activations are passed through the function.

- ***Pooling Layers (MaxPooling):*** The layers of pool are employed in order to cut the spatial size of the feature maps, so as to minimize the computational cost as well as the possibility of overfitting. The most widespread pooling operation is MaxPooling that automatically chooses the top value of a small area (e.g. 2×2 or 3×3) of the feature map. Such operation contains merely the most significant features of the picture and therefore minimises the size of feature maps.
- ***Fully Connected Layers (FC):*** The high-level features are flattened into a one-dimensional vector after going through a few convolutional layers, pooling layers followed by fully connected layers. These layers do the final predictions at the basis of the features that are extracted by the previous layers.

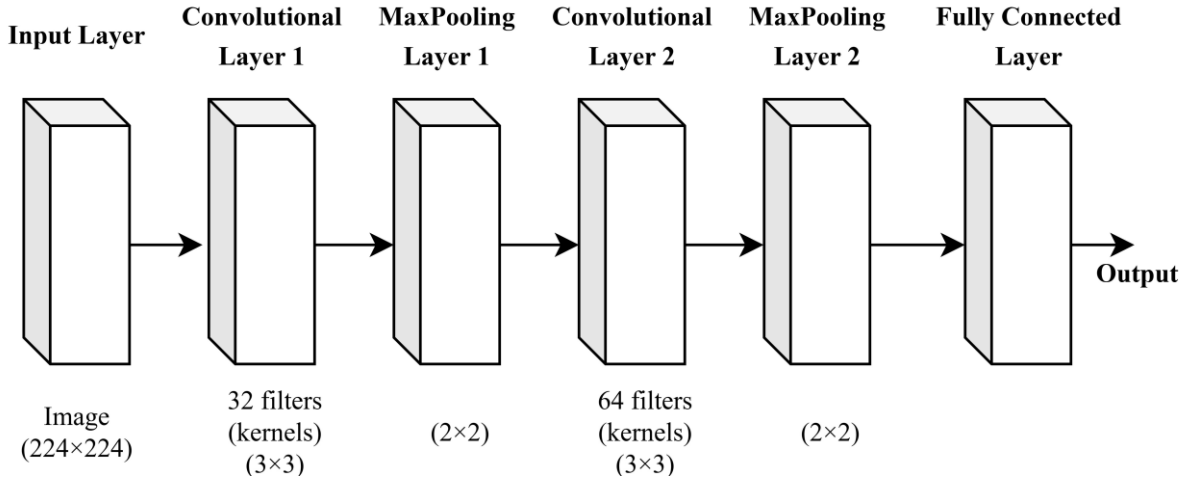

Figure 2: CNN Architecture for Soybean Leaf Disease Detection

The architecture of the Convolutional Neural Network (CNN) adopted to identify the presence of soybean leaf disease in leaves is shown in Figure 2. The network has an input image of size 224×224 pixels, which is followed by two convolutional layers where the features are extracted by using 32 and 64 filters respectively. The convolutional layer is successively followed by the max-pooling layer in order to minimize the space dimensions. Lastly, the features are given over a fully connected layer to make a prediction and the recognition is the predicted type of disease.

3.4 Combined Feature Vector

The features extracted using both HOG and CNN are combined and made into a combined feature vector. This compound feature vector combines low-level texture features with high-level learned features from the CNNs and represent aspects of the visual features of the soybean leaf in a complex way.

The formula of the combined feature vector in mathematical form is:

$$F_{combined} = [F_{HOG}, F_{CNN}] \tag{9}$$

Where:

- F_{HOG} = the feature vector formed by taking as input the HOG descriptor.
- F_{CNN} = the feature vector using the CNN.

This combined feature vector is used as input to the classification step.

3.5 Training the Classifier

The concatenation of feature vectors is split into training and testing data sets. The training set is used for training random forest classifier. Random Forest is a type of ensemble learning method which constructs several decision trees using random subsets of features. Each decision tree makes its own classification for healthy as well as diseased leaves and the final classification is reached with a majority vote of all trees.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

The prediction for the input feature vector x by each decision tree $T_j(x)$ is used to make the final prediction as follows:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_T(x)\} \tag{10}$$

where T is the total number of trees in the forest.

- ***Training Process:*** Each tree is trained on random subset of the data, based on Gini index or entropy, the best and optimal split is determined at each node of the tree.
- ***Gini Index:***

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{11}$$

where p_i is the probability of class i at node t, and C is the total number of classes.

3.6 Optimizing the Classifier Using Sine Cosine Algorithm (SCA)

The selection of the Sine Cosine Algorithm (SCA) in its role in optimization of the Random Forest (RF) classifier is pivotal in maximizing the classification accuracy. Although more basic optimization methods e.g. grid search or random search are sometimes employed to optimize hyperparameters, they may prove to be both time-consuming and inefficient, especially when dealing with large datasets. Grid search follows exhaustive search of all possible combinations of hyperparameters making it a time-intensive operation whereas random search does not give accurate selection of best values of all hyperparameters.

SCA, however, is a metaheuristic optimization algorithm, which models the exploration process, and the exploitation process of exploration procedure with the aid of sines and cosines functions. This algorithm has a number of strengths compared to simpler algorithms: it is computational, does not require exhaustive search, and has been shown to be useful in evading local optima, which is problematic with traditional optimization algorithms. The Random Forest classifier can be optimized better using SCA, which has the benefit of significantly better generalization at a lower cost of computation. Moreover, SCA is exceptionally suited to the dynamic adjustment of the exploration amplitude in the course of optimization technique due to the argument of non-linearity or multifaceted connection of needful traits with regard to the soybean leaf disease data.

For the SCA algorithm, the guidelines are:

- ***Sine Function for Exploration:***

$$S(t) = A(t) \cdot \sin(t) \tag{12}$$

- ***Cosine Function for Exploitation:***

$$C(t) = A(t) \cdot \cos(t) \tag{13}$$

where $A(t)$ is the amplitude, which diminishes over the time to make it easy for us to converge.

By optimizing the hyperparameters of Random Forest classifier, the SCA improves the performance of the classifier by decreasing the classification error. Following is the pseudo-code for the SCA algorithm.

1. Initialize a population of solutions randomly
2. Evaluate the fitness of each solution
3. While stopping criteria (max iterations or convergence) are not met:
    a. For each solution:
        i. Update its position using the sine and cosine functions
        ii. Generate random values to control the update step
    b. Evaluate the new positions (solutions)
    c. Update the best solution if a better one is found
4. Return the best solution found

3.7 Testing the Classifier

The trained and optimized model Random Forest is then tested on the unseen testing dataset. The classifier takes the feature vectors into consideration and predicts whether the leaf is diseased or healthy. During the testing process, the following metrics are used to measure the performance of the model:

- *Accuracy:*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

(14)

where: TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

- *Precision:*

$$Precision = \frac{TP}{TP + FP}$$

(15)

- *Recall / Sensitivity:*

$$Sensitivity = \frac{TP}{TP + FN}$$

(16)

- *F1-Score:*

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

(17)

3.8 Classification of Soybean Leaf Diseases

After the classifier is trained and optimized, the last step in the methodology would be the classification of the disease. In this step, the model applies the trained classifier Random Forest (RF) to predict the class (healthy or diseased) of a given soybean leaf. The process of disease classification includes taking the combined feature vector (based on combining texture features with the help of HOG and deep features with the help of CNN) and utilizing them to go through the trained classifier. The classifier then sends out a prediction of the disease on behalf of the leaf, which belongs to one of the categories defined previously, such as Asian Soybean Rust, Frogeye Leaf Spot or healthy.
Mathematically, the classification will be expressed as follows:

$$\hat{y} = RF(F_{combined})$$

(18)

Where:

- $\hat{y}$ mapped predicted label of the disease that reflects to input leaf image.
- RF stands for the Random Forest Classifier.
- $F_{combined}$ is representing the combined feature vector by merging texture features from HOG and deep features from CNN.

IV. Results and Discussion

All experiments in this research were developed and implemented with MATLAB as the main tool of the disease detection model. To execute the experiments, the hardware was set up with an Intel Core i7 processor, 16 GB of RAM, and a NVIDIA GTX 1660 Ti Gradient card, which could assist in fastening the computations of the deep learning. In the training and evaluation of the model, the dataset was divided into 70 percent and 30 percent training and testing respectively, which would give a good balance in balancing the model. The model has been trained under different techniques of deep learning including Convolutional Neural Networks (CNN) and data augmentation in order to enhance the structural strength and generalization of the model to different local environmental conditions.

4.1 Simulation Parameters

Table 1: Simulation Parameters

Parameter Name		Value
CNN	Number of Layers	5 (Convolutional + Pooling + Fully Connected)
	Kernel Size (Convolutional Layers)	3×3
	Activation Function	ReLU (Rectified Linear Unit)
	Pooling Type	MaxPooling
	Epochs	150
	Learning Rate	0.001
	Batch Size	32
Random Forest	Number of Trees	100
	Maximum Depth	10
	Minimum Samples Split	2
	Criterion	Gini Index
SCA	Population Size	50
	Max Iterations	1000

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

	Exploration Factor (Amplitude)	Decreases over iterations (from 1 to 0)
	Convergence Tolerance	0.001

4.2 Results

The results of the MATLAB based simulation experiments are presented in this section using the proposed methodology for soybean leaf disease detection. The simulations were conducted on a data set of soybean leaf images and the detection system was tested for its performance using different measures. The performance of the classifier, optimization effectiveness of SCA algorithm, and model accuracy of various soybean leaf diseases were presented in the following subsections.

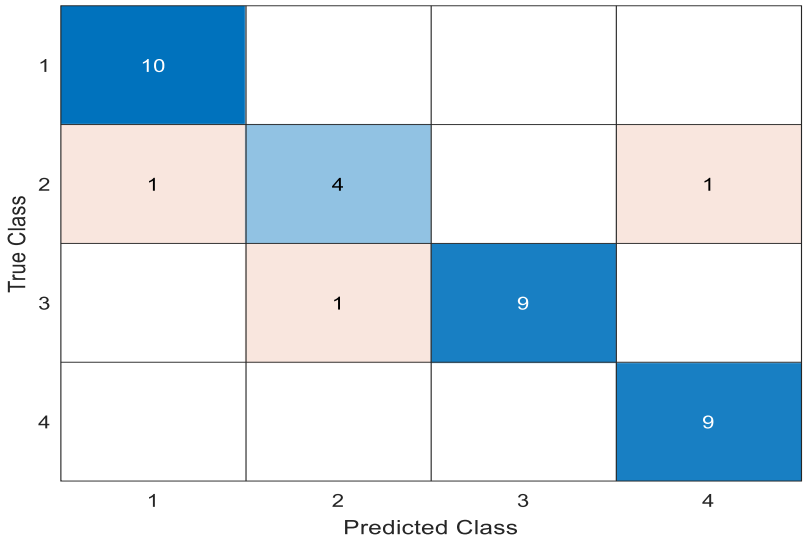

Figure 3: Confusion Matrix Output for Sine Cosine-Optimized Random Forest Classifier based Soybean Leaf Disease Detection

Figure 3 shows the confusion matrix output result of Sine Cosine-Optimized Random Forest Classifier in the detection of soybean leaf diseases. The performance of the classifier is shown in the following matrix for four predicted classes (1, 2, 3, 4) with respect to the true classes. The diagonal elements, the correct predictions, show guidelines high values for class 1 (10), class 3 (9) and class 4 (9), which displayed that these classes were classified correctly. However, there are certain misclassifications, for example in the off-diagonal elements. For instance, class 2 has 1 misclassification predicted as class 1, 4 as class 3 and 1 as class 4, so some overlap of classes is evident. Overall, the effectiveness of the classifier is observed to be good in terms of the identification of the soybean leaf diseases with some mis classifications that can be handled for further improvement of the classifier.

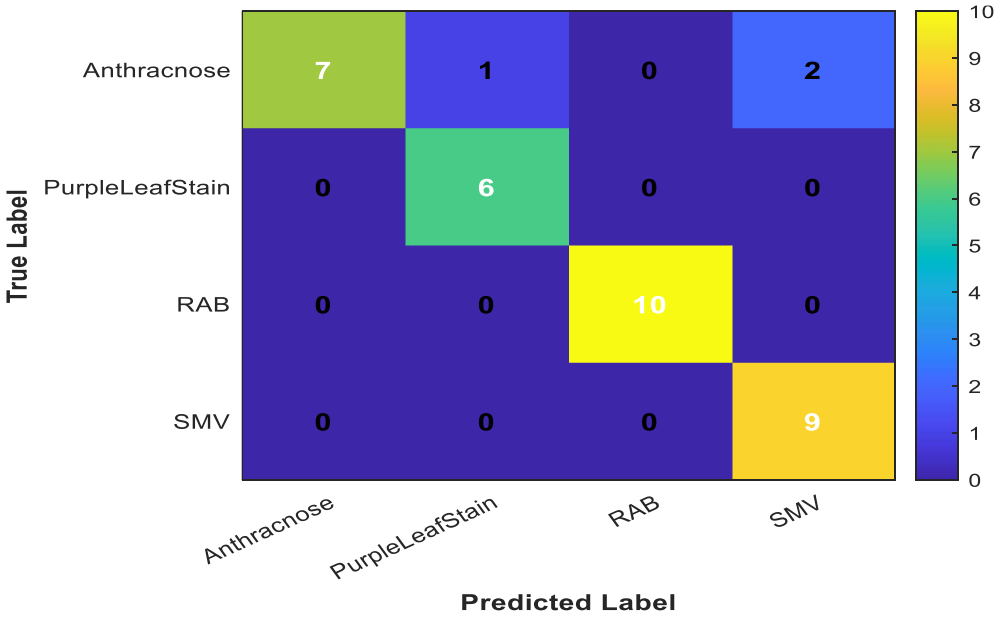

Figure 4: Confusion Matrix output for Disease Classification

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Figure 4 shows the confusion matrix output for the classification of the disease for the detection of soybean leaves. The matrix shows the performance of the classifier over four predicted classes which are Anthracnose, Purple Leaf Stain, RAB and SMV. The diagonal values which signify correct predictions reveal 7 for Anthracnose, 6 for Purple Leaf Stain, 10 for RAB and 9 for SMV, which indicate that these diseases are correctly predicted. However, there are some misclassifications, as one can see in the off-diagonal elements. For example, 1 instance of Anthracnose was predicted to be Purple Leaf Stain and 2 instances of Anthracnose were predicted to be SMV. In addition, 1 instance of RAB was predicted to be SMV. These misclassifications illustrate areas where the accuracy of the classifier might be further improved, however, overall, the system is shown with strong accuracy in its performance.

Figure 5 shows the convergence plot for Sine Cosine Optimization (SCA). The plot helps to determine the best accuracy of the optimization process iterations. As at first the accuracy is stable at around 88.2% until the 7th iteration, a huge improvement is noticed and thus leading to a sharp improvement in accuracy. The accuracy levels off at around 88.7% after the jump. This sharp improvement in accuracy shows how effective the SCA is to optimize the classifier, and it was found that the convergence is accomplished already towards the later iterations.

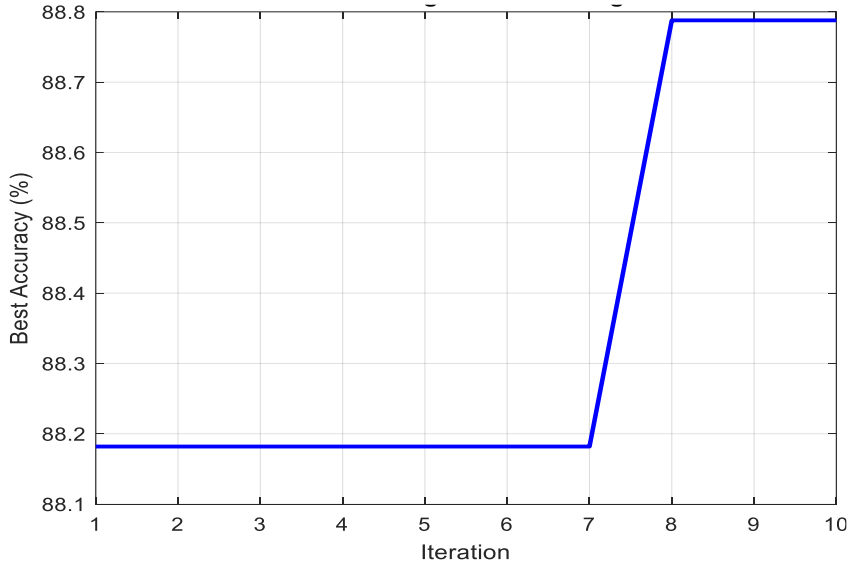

Figure 5: Convergence Plot for Sine Cosine Optimization

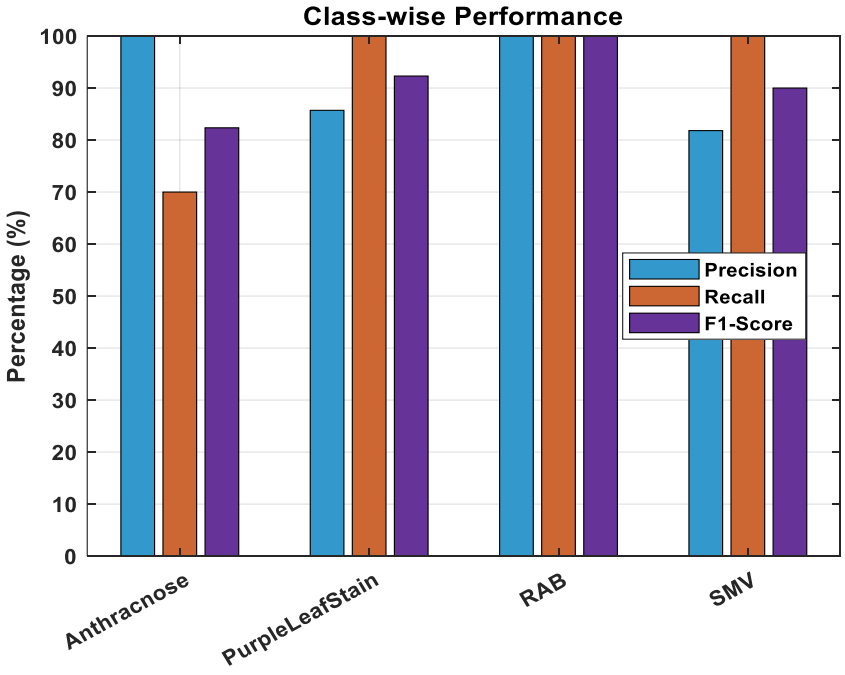

Figure 6: Bar Chart Comparison for Class-wise Performance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Figure 6 is a bar chart comparison for class-wise performance of the classifier that displays the Precision, Recall, and F1-Score for four different types of leaf diseases of soybean: Anthracnose, PurpleLeafStain, RAB and SMV. The performance measures are outlined in the chart as percentages, which are given using three bars for Precision (blue), Recall (orange) and F1-Score (purple) in each row. Precision and Recall values are quite high for all the classes, with most diseases resulting in values approaching and surpassing 90%. For instance, the Precision and Recall scores of Anthracnose and SMV are above 90%, while the F1-Score scores of these classes are also high, from the other point of view they show that there is a balanced performance. PurpleLeafStain and RAB have a bit lower precision and recall than the other diseases but have good performance, and F1-Scores can be seen that are close to 90% for the three classes, showing the reliable performance of the classifier in the dataset.

Table 2: Performance Evaluation Results

Parameter Name	Value
Accuracy	91.43%
Precision	90.23%
Recall	89.17%
F1-Score	89.36%
ELAPSED TIME	180.079 seconds

Table 2 shows the results of performance evaluation of the Sine Cosine Optimized Random Forest Classifier for soybean leaf diseases classification. The classifier attained an overall accuracy of 91.43% which means that the model was able to accurately predict the correct class for a large proportion of the total number of instances in the dataset. The accuracy of 90.23% indicates how often the model predicted actual positive cases out of all the positive predictions the model made, implying that there is only a small number of false positives. For actual positive cases, the model can remember about 89.17%, which indicates that the model has good discrimination skills, however, there is a little loss in detecting some positive cases. The F1-Score of 89.36% particularly is a balanced measure of both precision and recall and provides a representation of how well the classifier is doing when faced with the exchange between false positive classification and false negatives. The computational time of the whole evaluation process was 180.079 seconds and showed the computational efficiency of the model during its execution, which makes it an appropriate model for real-time disease detection applications in large-scale agricultural environment. These results reflect the overall good performance of the model in the detection of soybean leaf diseases.

Table 3: Comparative Analysis of Proposed Research Work with Previous Research Works

Ref. No.	Methodology Used	Accuracy	Precision	F1-Score
[5]	Improved ConvNeXt Deep Learning Model	85.42%	--	--
[12]	Faster R-CNN (MF^3 R-CNN) Model		83.34%	--
[13]	Least Absolute Shrinkage and Selection Operator (LASSO) regression	87.5%	--	--
Proposed Approach	**Deep Learning and SCA-Optimized Random Forest**	**91.43%**	**90.23%**	**89.36%**

Table 3 shows a comparative analysis of proposed research with the previous research texts in the field of soybean leaf disease detection. The table emphasises on methodologies used, accuracy, precision and F1-Score of each of the studies. The proposed approach in this research, which is a combination of deep learning and SCA-optimized Random Forest, is better than other methods with the highest accuracy (91.43%), precision (90.23%) and F1-score (89.36%). In comparison, the [5] ConvNeXt Deep Learning Model attained an accuracy of 85.42%, but no precision and F1 score values were reported. The [12] Faster R-CNN (MF^3 R-CNN) Model The Faster R-CNN has an accuracy of 83.34% without any additional performance metrics available. The [13] LASSO regression model performed by 87.5% accuracy, however the precision and F1 scores values were not mentioned as in the other studies. This comparison clearly demonstrates the proposed methodology to deliver better classification performance in comparison with the existing approaches in terms of both accuracy and balance between precision and recall demonstrating the effectiveness of this methodology for large-scale disease detection in soybean crops.

4.3 Discussion

The results of this research show the improvement that was made in depth on the detection of soybean leaf diseases with the use of deep learning technique and machine learning. By settings and optimizations of the Random Forest classification through the Sine Cosine Algorithm (SCA), the classification accuracy of the random forest is very high, which is 91.43%, which is a very good improvement compared to the traditional classification method. The combinations of modern feature extraction algorithms like HOG and CNN have further facilitated this success, which makes it possible to detect diseases more accurately by capturing

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

important visual features of the soybean leaves. The capacity of the system to deal with large data efficiently and its real-time applicability are of great value for farmers, enabling them to intervene against crop damage in time. In addition, the system meets some of the challenges typically experienced in agricultural diseases detection, including noise, environmental variation, and high amounts of data. However, while the model performs well in a controlled condition, there is still a need for further optimization to improve robustness and adaptability in diverse agricultural situations to make them have a consistent performance under different field conditions.

4.4 Limitations

- While the model shows high precision in lab conditions, it cannot be assumed to be the same in reality due to environmental variations like changing kinds of light, weather conditions and so on that might interfere with image quality that might affect prestress classification accuracy.
- The besides computational needs, especially the optimization process, of the system can be challenging for implementation on devices with low resources in rural settings.
- Although the dataset used is extensive it might not represent the immense variation of soybean leaf diseases among different geographical regions, which will potentially have the drawback of reducing the generalization of the model in certain geographical regions.

V. CONCLUSION

The paper introduces a machine learning and deep learning-based automated system in detection of soybean leaf diseases. The system utilizes features extraction systems such as HOG and CNN, which have been optimized using the Sine Cosine Algorithm (SCA) on the Random Forest Classifier with a classification accuracy of 91.43%. The approach is superior compared to the conventional methods and provides a cost-efficient scaling of real-time disease detection amongst mass scale agricultural applications. The system is capable of managing image noise, environmental changes, and massive volumes of data and hence it could be a hopeful tool to be used to modernize soybean farming. Nevertheless, it requires even more advancement in the real-life application, particularly in following the shifts in the environment. Further research will be aimed at data acquisition and transfer learning method as well as field validation, which is based on the IoT to increase the robustness of the system and extend its applicability to other crops and geographical areas.

REFERENCES

[1] Goshika, S., Meksem, K., Ahmed, K.R. and Lakhssassi, N., 2023. Deep learning model for classifying and evaluating soybean leaf disease damage. International Journal of Molecular Sciences, 25(1), p.106.
[2] Pranta, A.S.U.K., Fardin, H., Debnath, J., Hossain, A., Sakib, A.H., Ahmed, M.R., Haque, R., Reza, A.W. and Dewan, M.A.A., 2025. A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification. Computers, 14(5), p.197.
[3] Adimas, A.K., Mekonen, M.Z., Assegie, T.A., Singh, H.K., Mazumdar, I., Gupta, S.K., Salau, A.O. and Tin, T.T., 2025. Soybean leaf disease detection and classification using deep learning approach. Bulletin of Electrical Engineering and Informatics, 14(4), pp.2697-2704.
[4] Yu, M., Ma, X., Guan, H., Liu, M. and Zhang, T., 2022. A recognition method of soybean leaf diseases based on an improved deep learning model. Frontiers in plant science, 13, p.878834.
[5] Wu, Q., Ma, X., Liu, H., Bi, C., Yu, H., Liang, M., Zhang, J., Li, Q., Tang, Y. and Ye, G., 2023. A classification method for soybean leaf diseases based on an improved ConvNeXt model. Scientific Reports, 13(1), p.19141.
[6] Alnuaim, A., Altheneyan, A. and AlZubi, A.A., 2025. Early Leaf Disease Detection of Soybean Plants using Convolution Neural Network Algorithm. Legume Research: An International Journal, 48(6).
[7] Sharma, V., Tripathi, A.K., Mittal, H. and Nkenyereye, L., 2025. SoyaTrans: A novel transformer model for fine-grained visual classification of soybean leaf disease diagnosis. Expert Systems with Applications, 260, p.125385.
[8] Yin, J., Li, W., Shen, J., Zhou, C., Li, S., Suo, J., Yang, J., Jia, R. and Lv, C., 2025. A Diffusion-Based Detection Model for Accurate Soybean Disease Identification in Smart Agricultural Environments. Plants, 14(5), p.675.
[9] Feng, J., Wu, S., Mu, R., Xu, H., Zhai, Z. and Hu, B., 2025. Stoma Detection in Soybean Leaves and Rust Resistance Analysis. Plants, 14(19), p.2994.
[10] Bhavani, R., 2025. Detection of Leaf Diseases in Soybean Plant using Autoencoder and Multinomial Logistic Regression. Legume Research: An International Journal, 48(5).
[11] Annrose, J., Rufus, N.H.A., Rex, C.E.S. and Immanuel, D.G., 2022. A cloud-based platform for soybean plant disease classification using archimedes optimization based hybrid deep learning model. Wireless Personal Communications, 122(4), pp.2995-3017.
[12] Zhang, K., Wu, Q. and Chen, Y., 2021. Detecting soybean leaf disease from synthetic image using multi-feature fusion faster R-CNN. Computers and Electronics in Agriculture, 183, p.106064.
[13] Zhang, H. and Zhang, H.H., 2023. Classification of soybean diseases by logistic regression with LASSO using phenotypic and environmental features. Journal of High School Science, 7(3).
[14] Song, H., Huang, Y., Han, T., Xu, S. and Liu, Q., 2025. A cell P system with membrane division and dissolution rules for soybean leaf disease recognition. Plant Methods, 21(1), p.39.
[15] Chen, C., Lu, X., He, L., Xu, R., Yang, Y. and Qiu, J., 2025. Research on soybean leaf disease recognition in natural environment based on improved Yolov8. Frontiers in Plant Science, 16, p.1523633.
[16] Jahin, M.A., Shahriar, S., Mridha, M.F., Hossen, M.J. and Dey, N., 2025. Soybean Disease Detection via Interpretable Hybrid CNN-GNN: Integrating MobileNetV2 and GraphSAGE with Cross-Modal Attention. arXiv preprint arXiv:2503.01284.
[17] Thakur, P.S., Khanna, P., Sheorey, T. and Ojha, A., 2022. Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. arXiv preprint arXiv:2207.07919.
[18] Bhageerathi, T., Anagha, M. and Pushpa, T.S., 2024. Comparative Analysis of Deep Learning Models for Plant Disease Detection. ResearchGate. https://www.researchgate.net/publication/376274780
[19] Soybean (Large Dataset). UCI Machine Learning Repository. Available online at: https://archive.ics.uci.edu/dataset/90/soybean+large