RESEARCH ARTICLE Legume Research- An International Journal, Volume 48 Issue 5: 876-884 (May 2025)

Detection of Leaf Diseases in Soybean Plant using Autoencoder and Multinomial Logistic Regression

R. Bhavani[1]

10.18805/LR-5461

ABSTRACT

Background: Soybean is one of the important leguminous crops grown mainly in the middle states of India. Soybean plants are prone to leaf diseases like spot and bacterial blight. Early detection of such diseases is an important task for the farmers to avoid loss in production. In this background, deep learning techniques are used for identifying the diseases in leaves of Soybean plants.

Methods: This study investigates the utilization of convolutional autoencoder in extracting features from the lesion leaf images. Images are preprocessed and converted to latent space features using convolutional autoencoder. Multinomial logistic regression model is employed over the extracted features to find the type of disease in Soybean plants.

Result: The experimental result shows that convolutional autoencoder model along with multinomial logistic regression model achieves an average accuracy of 92% in disease identification. In addition to providing disease identification, the ideas in this paper provide support for food security through the application of deep learning techniques.

Key words: Convolutional autoencoder, Leaf disease detection, Multinomial logistic regression, Soybean leaf disease.

INTRODUCTION

Soybean (*Glycine max*) is one of the most broadly grown legume plant in Madhya Pradesh and Maharashtra, India for its oil and protein products (Jianing *et al.,* 2022). The yield of the plant greatly depends on the health of the plant which in turn depends on the resistance of the plant for leaf disease. The common leaf diseases in soybean plants include soybean leaf spots and soybean rust. The early detection of these diseases is important for better productivity. In early days these disease identification were done manually and accurate identification of these disease types was a time consuming process (Wu *et al.,* 2023).

The rapid developments of image processing and machine learning techniques have led to algorithms for automatic disease detection from the leaf images of the diseased plants. Image processing aims at deriving useful information by performing many operations on the digital images (Saradhambal *et al.,* 2018). Image processing finds its application in the various research fields such as cyber security, telemedicine, agriculture, *etc*. (Prabaharan *et al.,* 2020). Machine learning has its application in the field of agriculture to design automatic harvesting machines; estimate production; managing irrigation needs; pest and weed control activities, *etc*. (Yao *et al.,* 2023).

Machine learning algorithms used for feature extraction and leaf disease classification are discussed in the rest of this section. Leaf images were segmented using partition based clustering algorithm namely, K-means clustering algorithm to identify the lesion part of the leaf and then colour and shape features are extracted from it. Machine learning techniques like decision tree (DT), Support vector machines (SVM) and K-Nearest neighbors (K-NN) are applied to classify the leaf diseases (Nandhini and Bhavani, 2020). Leaf images in RGB, HSV and Lab* colour space were used to extract texture features such as Gray-Level Run-Length Matrix and the Gray-Level Occurrence Matrix from the chickpea plant leaf images. Multi-class classification models such as K-NN, SVM and Neural Networks were used to classify leaf diseases. Their proposed model works well in identifying fusarium wilt of chickpea leaves (Hayit *et al.,* 2024). Colour features like HSV features were obtained from segmented images of lesion leaves to train the artificial neural network (ANN) to distinguish the healthy and diseased cotton leaf samples (Ranjan *et al.,* 2015). Texture features such as contrast, energy, homogeneity, correlation and smoothness along with region shaped shape features were extracted from leaf images and Adaptive Neuro Fuzzy Inference system was used for disease identification (Nandhini and Srisathya, 2021). Statistical features like colour co-occurrence matrix and shape features using blob analysis were extracted from segmented leaf images and disease classification was done using SVM (Kappali *et al.,* 2024). These five works in the literature, extracted colour, shape or texture features from the segmented plant leaves and applied SVM, K-NN, ANN, ANFIS, RF, LR or ensemble

[1]Department of Computer Science and Engineeirng, Government College of Engineering, Srirangam, Trichy-620 012, Tamil Nadu, India.

Corresponding Author: R. Bhavani, Department of Computer Science and Engineeirng, Government College of Engineering, Srirangam, Trichy-620 012, Tamil Nadu, India.
Email: bhavanirajasekar@gmail.com
ORCID: 0000-0002-8849-8311

How to cite this article: Bhavani, R. (2025). Detection of Leaf Diseases in Soybean Plant using Autoencoder and Multinomial Logistic Regression. Legume Research. 48(5): 876-884. doi: 10.18805/LR-5461.

Submitted: 18-12-2024 **Accepted:** 08-03-2025 **Online**: 19-04-2025

classification techniques for identifying leaf diseases and achieved average accuracy in the range of 80% to 90%. It is found in literature that autoencoder, a deep neural model, is used for image denoising, image reconstruction, image compression, feature extraction, *etc*. (Li et al, 2023). Stacked denoise autoencoder (SDAE) was used to extract features from hyperspectral images and logistic regression (LR) approach was employed for classification (Xing *et al.,* 2016). Convolutional autoencoder was used in feature extraction from Optical Emission Spectroscopy (OES) data samples and Support Vector Regression (SVR) machine was for predicting final *etc*h rate (Maggipinto *et al.,* 2018). A multitask learning framework based on siamese network and autoencoder was developed for classification of hyperspectral images (Miao *et al.,* 2019). Autoencoders were applied in converting MRI image to feature vectors (Chen *et al.,* 2024).

With the advancement in deep learning techniques, leaf disease classification was done by applying Convolutional Neural Network (CNN) models. Leaf diseases in Faba bean plants were identified using a CNN model (Jeong and Na, 2024). VGG16 pretrained model was enhanced with a stack of one convolutional layer, one pooling layer and one fully connected layer in order to detect and classify Wilting in Soybean crop (Na and Na, 2024). Foliar diseases in black gram plants were detected using CNN (Kalpana *et al.,* 2023). A customized CNN model was designed to identify diseases in leaves of tomato plants. The results obtained were compared with the pre-trained models of VGG16, InceptionV3 and MobileNet (Agarwal *et al.,* 2020). A web based application to identify fungal and bacterial diseases in potato leaves was built using image segmentation techniques and CNN model (Shukla and Sathiya, 2022). All these methods used a customized CNN model or improvised the existing CNN model to classify leaf disease in various types of plants. On an average these models achieved accuracy in the range of 85% to 90%.

With the advent of autoencoders for feature extraction in various fields like hyperspectral image classification, OES classification, MRI image retrieval, *etc*., this paper aims to develop a classification model using convolutional autoencoder and Multinomial logistic regression. Convolutional autoencoder is used for extracting features from the Soybean leaf images and Multinomial logistic regression model is applied in classifying the type of disease from the extracted features.

MATERIALS AND METHODS

The details about dataset collection, pre-processing, feature extraction, foliar disease identification and performance metrics to evaluate the model are discussed

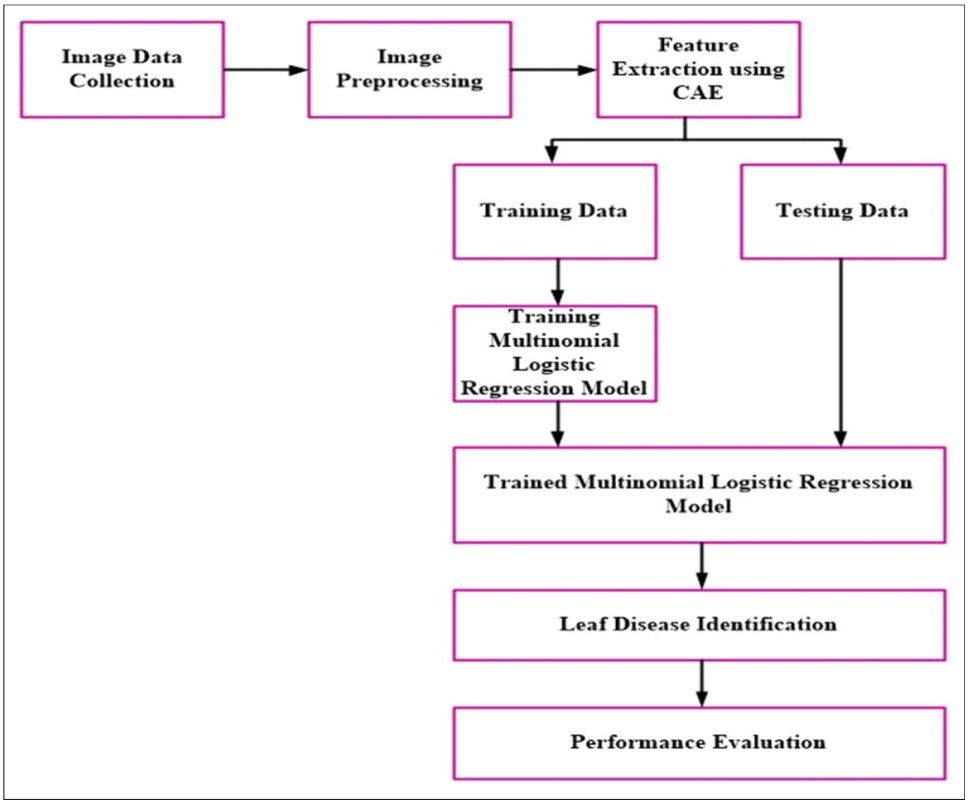


Fig 1: Overall working of the proposed model.

in this section. This experiment was carried out during July to November 2024 at Government College of engineering Srirangam, Trichy. Fig 1, shows the overall flow of data of the proposed model.

Dataset collection

The first step in constructing a deep learning model is collecting appropriate data. This step is very important for deep learning models as they provide the foundation for training the models and help them make accurate predictions. The dataset containing five classes of Soybean leaf disease like healthy, vein necrosis, dry leaf, septoria brown spot and bacteria leaf blight (Kotwal *et al.*, 2024) with 288, 138, 230, 284 and 226 images respectively was downloaded. Sample leaves for each class of Soybean leaf disease is shown in Fig 2.

Image pre-processing

In this process, a set of techniques are employed to prepare the images suitable for processing by the deep learning model. This step enhances the quality of the image before they are further analysed and processed by deep learning algorithms. The pre-processing steps done for this work are resizing, normalization and data augmentation. Leaf images are first resized to a fixed size of 254 × 254 pixels, then normalized using z-score normalization. Autoencoder model requires a huge training data to prevent overfitting. Using techniques such as rotation, flipping and brightness adjustments, the normalized images are amplified to get a diverse dataset.

Feature extraction

In this step, raw images are transformed into numerical features without compromising the information in the original image. Although there are different techniques for feature extraction like colour features, texture features and shape features (Alsmadi 2020), one remarkable approach is the use of autoencoders (Cambuí *et al.,* 2021).

Autoencoders are neural network models that are capable of reconstructing the input data. An autoencoder has three major components such as an encoder, a bottleneck and a decoder as shown in Fig 3. The encoder attempts to compress the resized input image into latent representation and the decoder attempts to recreate the image from the compressed data. The intermediate part between the encoder and the decoder is the bottleneck which contains compressed data of the input image. Initially, the autoencoder model is trained to reconstruct the given image. After successful training, the decoder is removed and the latent space representation in the bottleneck serves as the features for the given image. Thus, the encoder part of autoencoder could be used for feature extraction from raw data and this compressed data is used in training the machine learning model.

Convolutional autoencoder

Convolutional autoencoder is a type of autoencoder that is more suitable for processing images (Fig 4). It is built using convolutional neural networks. convolutional autoencoder helps in analysing image data by capturing the spatial

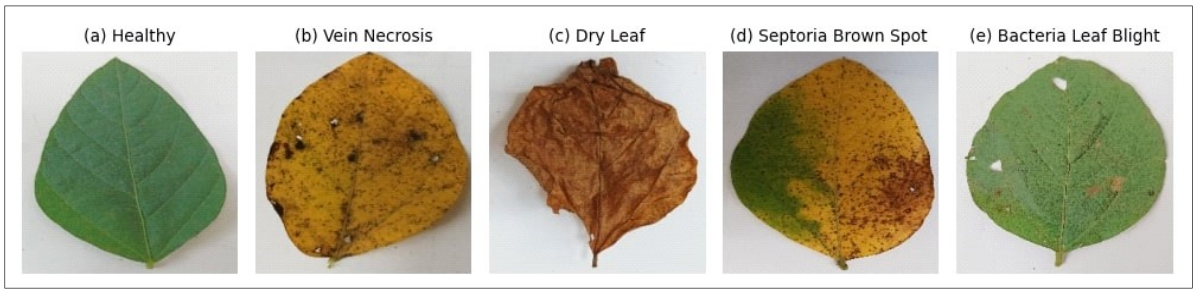


Fig 2: Sample leaf images.

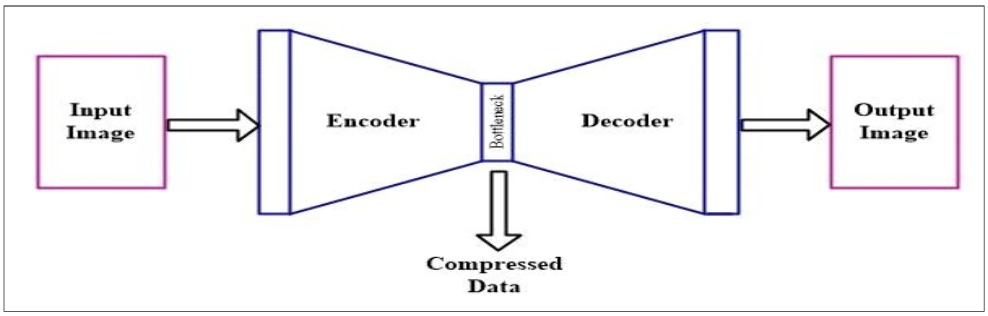


Fig 3: Autoencoder model.

relationships between the pixels in it (Polic *et al.,* 2019). The encoder part of CAE is constructed using a series of convolutional layer each followed by max-pooling layer. The convolutional layer works by sliding the filters over the pixels of the image to perform pixel wise multiplication and summing up together to form a single pixel. After convolution operation, the number of pixels in the intermediate feature maps is further reduced by applying pooling operation. Max-pooling operation takes the maximum value inside the window of specified size, slides over the feature map and obtains a feature map of reduced size. The bottleneck part of CAE is formed by flattening the convolutional layer at the end of the encoder such that the input image is converted to the latent space with fewer dimensions. The decoder part of CAE works as the opposite of the encoder part. It first performs upsampling of feature maps and then applies the convolution operation to reconstruct the original image. This paper aims at extracting features from encoder part of the trained autoencoders which is further used to train the multinomial logistic regression model to identify foliar diseases.

Architecture of the proposed CAE

The first part of the proposed work is to build a convolutional autoencoder for reconstructing the lesion images of soybean leaves. After training the CAE model, the decoder part is discarded and the encoder part is used for extracting features from leaf images. The architecture of the CAE model is this experiment given in Table (1). The encoder part has six convolutional layers with max-pooling layer following each of them. The filter size of 3×3 and stride of 1 is used to perform convolution operation. In the proposed architecture, relu activation function is used the convolution layers and is given in equation (1).

$$f(x) = \begin{cases} 0 \text{ for } x \leq 0 \\ x \text{ for } x > 0 \end{cases} \quad ...(1)$$

The max-pooling operation pools data with 2×2 window size with stride of 2. The decoder part consists of six convolutional layers each followed by an up-sampling layer. The up-sampling operations works with an up-sampling factor of 2×2 and interpolation is done using the nearest neighbor. The last convolutional layer outputs the reconstructed image of size 254×254×3.

Model training

After pre-processing, the images are splitted into three groups; one for training; another for validation and the rest for testing the convolution autoencoder. They are splitted in the ratio of 80:10:10. That is, 80% in each category of the image set is taken for training the CAE model, 10% is used for validating the CAE model and another 10% is for testing the CAE model. The hyper parameter of the proposed CAE model is assigned with batch-size of 32 and optimization

Table 1: Proposed architecture of the CAE model.

Layer	Filter size	Output size
Input Layer	-	254×254×3
Conv2D	3×3	254×254×32
max-pool	2×2	127×127×32
Conv2D	3×3	125×125×64
max-pool	2×2	62×62×64
Conv2D	3×3	60×60×64
max-pool	2×2	30×30×64
Conv2D	3×3	28×28×64
max-pool	2×2	14×14×64
Conv2D	3×3	12×12×64
max-pool	2×2	6×6×64
Conv2D	3×3	4×4×64
max-pool	2×2	2×2×64
Conv2D	3×3	2×2×64
up-sample	2×2	4×4×64
Conv2D	3×3	6×6×64
up-sample	2×2	12×12×64
Conv2D	3×3	14×14×64
up-sample	2×2	28×28×64
Conv2D	3×3	30×30×64
up-sample	2×2	60×60×64
Conv2D	3×3	62×62×64
up-sample	2×2	125×125×64
Conv2D	3×3	127×127×32
up-sample	2×2	254×254×32
Conv2D (Output)	3×3	254×254×3

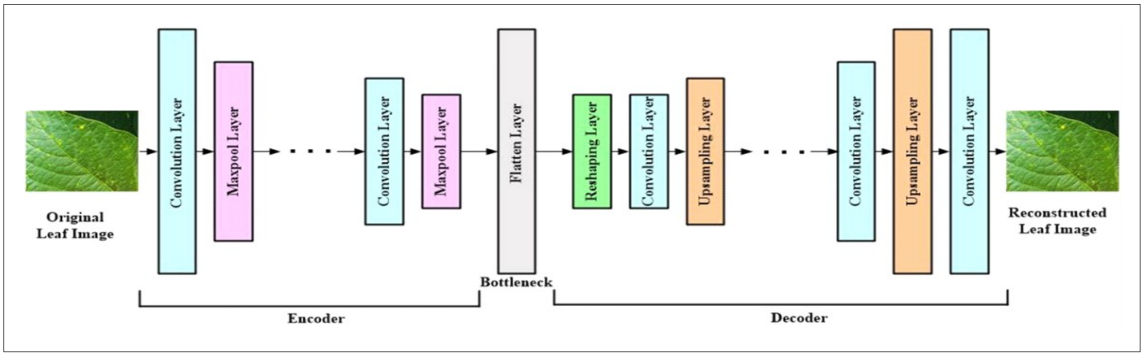


Fig 4: Convolutional autoencoder model.

with Adam optimizer and number of epochs for training with 50. The training performance is evaluated using Mean Square Error (MSE) and is given in equation (2).

$$MSG = \frac{\sum_{i=1}^{n} (f_i - y_i)^2}{n} \quad ...(2)$$

Where,
yi = Actual output.
fi = Output identified by the model.
n = Count of input images used for training.

After training the CAE model, the features are extracted using the encoder part of the CAE model. The output of the encoder is flattened to get the latent space representation of the lesion leaf image. The input size of 254×254×3 is reduced to features of size 256 (*i.e.*, 2×2×64).

Leaf disease Identification

After extracting features from lesion leaf images, the type of disease in soybean leaves is identified using multinomial logistic regression model. This type of regression model is applied on datasets containing more than two class labels (El-Habil, 2012). As the number of classes in our dataset is five, Multinomial Logistic Regression is chosen in our experiments. The functioning of the Multinomial Logistic Regression model is shown in Fig 5.
For a given instance

$$X = (x1, x2, ...xm)$$

multinomial logistic regression model first calculates the softmax score for each class using equation (3).

$$Zi = W^T X + b^i \quad ...(3)$$

such that $W \varepsilon R^{K \times m}$
Here k = Count of distinct classes in the dataset.
m = Size of features in the input data. In our experiments.
k=5.
m=256.

Next it computes the class probability of the given instance by applying softmax function (equation 4) over the softmax score such that the sum of all the class probabilities is equal to 1.

$$pi = \sigma_{softmax}(z_i) = \frac{\exp(zi)}{\sum_{j=i}^{K} \exp(zi)} \quad ...(4)$$

Finally the argmax operator (equation 5) returns the class label for the particular instance that maximizes the class probabilities.

$$y = \underset{i=1 \text{ to } k}{\text{argmax}} (p_k) \quad ...(5)$$

This model is trained by minimizing the cross entropy function using gradient descent technique.

Performance metrics

Multinomial logistic regression model is evaluated for its performance in identifying the class of the Soybean leaf disease using precision, recall, accuracy and F1-score metrics (Jiawei and Micheline, 2006).

Confusion matrix

The Table drawn with the count of correctly and wrongly classified data is the confusion matrix and its general model is shown in Table (2). The terms in the table are TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative). If the actually positive labelled sample is identified with positive label then it is counted in TP. If the actually negative labelled sample is identified with positive label then it is counted in FP. If the actually negative labelled sample is identified with negative label then it is counted in TN. If the actually positive labelled sample is identified with negative label then it is counted in FN.

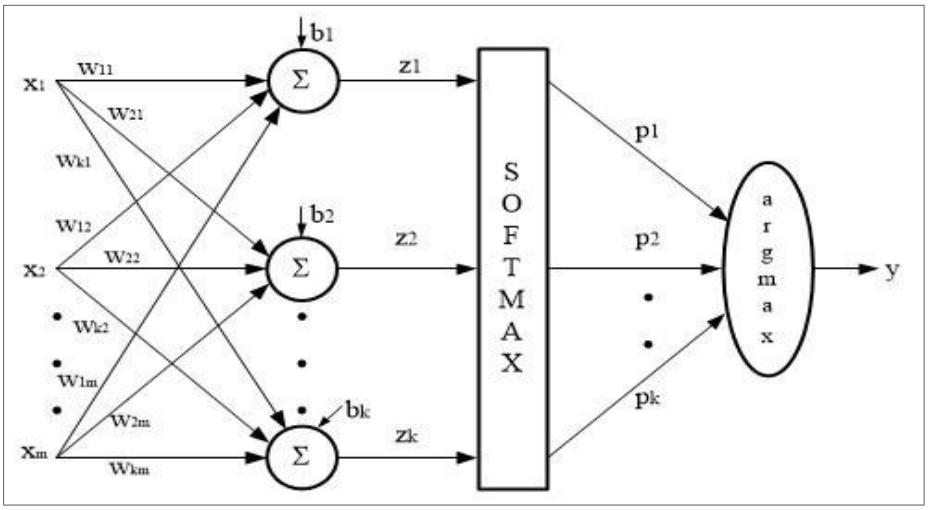


Fig 5: Multinomial logistic regression model.

Precision

Precision gives the quality of positive predictions of a machine learning model. It is calculated using equation (6).

$$\text{Precision} = \frac{TP}{TP + FP} \quad ...(6)$$

Recall

Recall gives the true positive rate of the classification model and is calculated using equation (7).

$$\text{Recall} = \frac{TP}{TP + FN} \quad ...(7)$$

Accuracy

It gives the overall performance of the model and is given by equation (8).

$$\text{Accuracy} = \frac{TP + TN}{P + N} \quad ...(8)$$

Where
P = Count of positive samples.
N = Count of negative samples used for testing.

F1-score

F1 score is the metric that uses the combined score of precision and recall and is given by equation (9).

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad ...(9)$$

RESULTS AND DISCUSSION

The performance of Convolutional autoencoder in feature extraction and the performance of Multinomial Logistic Regression in leaf disease identification are discussed in this section. All the experiments were done using the Python programming language. Libraries such as keras, numpy and opencv were used to build the deep learning models.

Performance of CAE in feature extraction

The encoder part of trained CAE outputs 256 features from the original image of size 254×254×3. The CAE model was trained in batches of 32 images, each of size 254×254×3. After 50 epochs of training the CAE model, the training loss obtained is 0.0826 (Fig 6) and the accuracy obtained is 92.32% (Fig 7). Further, the validation loss and validation accuracy is 0.0663 and 93.31% respectively. This result shows that the CAE model performs relatively similar during the training and validation phase. This signifies that the proposed CAE model performs well in reconstructing the lesion leaf image from the training data and could be used for extracting features.

Performance of multinomial logistic regression

The confusion matrix for the process of leaf disease identification using Multinomial Logistic Regression model is given in Fig (8). The labels of each class presented in the confusion matrix are, HY - Healthy leaf, VN -Vein Necrosis, DL - Dry leaf, SBS - Septoria brown spot, BLB - Bacterial Leaf Blight. As 10% of data is used for testing, there were 29, 14, 23, 28 and 22 leaf images tested in HY, VN, DL, SBS, BLB classes respectively. The performance metrics of the Multinomial Logistic Regression model obtained from confusion matrix is tabulated in Table (3). It is observed that, even though there are deviations in the

Table 2: Confusion matrix.

		Predicted labels	
		Positive	Negative
Actual labels	Positive	TP	FN
	Negative	FP	TN

Source: Jiawei and Micheline (2006).

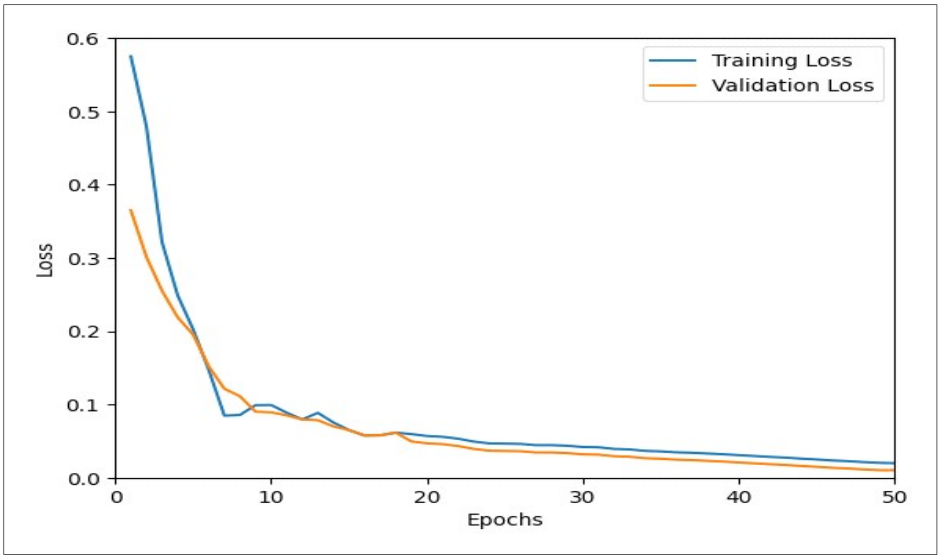


Fig 6: Training loss vs validation loss.

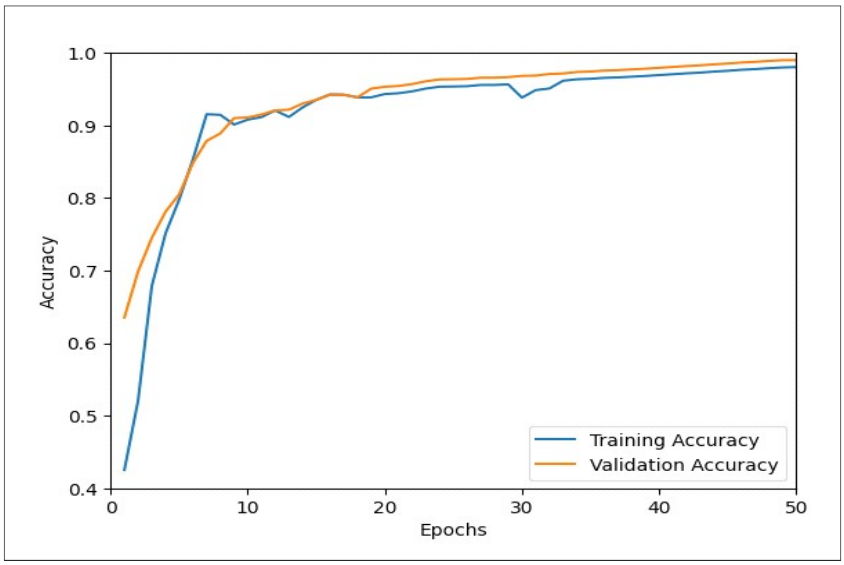


Fig 7: Training accuracy vs validation accuracy.

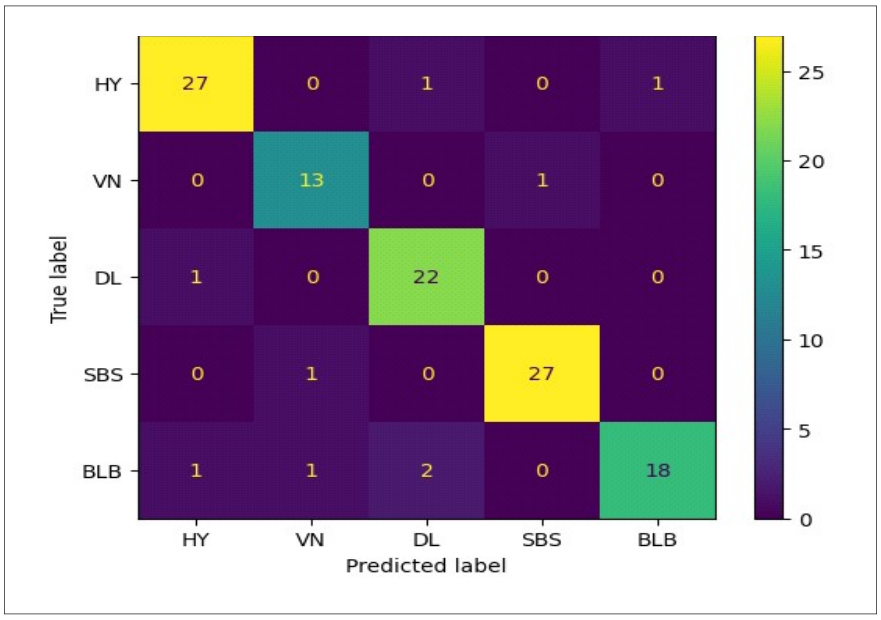


Fig 8: Confusion matrix for soybean leaf disease identification.

Table 3: Classification report of multinomial logistic regression.

Class label	Precision	Recall	F1-score
Healthy leaf	0.93	0.93	0.93
Vein necrosis	0.87	0.93	0.90
Dry leaf	0.88	0.96	0.92
Septoria brown spot	0.96	0.96	0.96
Bacterial leaf blight	0.95	0.82	0.88
Accuracy			0.92
Macro avg	0.92	0.92	0.92
Weighted avg	0.92	0.92	0.92

class wise observation of precision, recall and F1-score, Multinomial Logistic Regression model works well with the features learned from the CAE model in identifying the leaf disease of the Soybean plant with an overall accuracy of 92%.

In future, investigation is needed to gather soybean leaf images from diverse places, different climatic conditions and disease conditions. The model proposed in this paper could be improved further, by tuning the hyper parameters such as batch size, learning rate of Adam optimizer and epochs.

CONCLUSION

This current study examined the use of Convolution Autoencoder and Multinomial Logistic Regression in the context of Soybean leaf disease identification. Features from the soybean leaf images were extracted using Convolutional Autoencoder which is further used to identify the type of disease using Multinomial Logistic Regression. The model used in this study identifies five classes of Soybean leaf disease. The experimental results show that the Multinomial Logistic Regression model used in soybean leaf disease identification yields an overall accuracy of 92% from the features learned from the Convolutional Autoencoder model. This hybrid deep learning model could be used efficiently in detecting the leaf diseases of Soybean plants in the early stages of infection such that the production loss could be mitigated thus benefiting the farmers.

ACKNOWLEDGEMENT

None.

Disclaimer

The author assumes full responsibility for the information provided but disclaims any liability for any direct or indirect losses arising from the use of this content.

Conflict of interest

All authors declared that there is no conflict of interest.

REFERENCES

Agarwal, M., Singh, A., Arjaria, S., Sinha, A. and Gupta, S. (2020). ToLeD: Tomato leaf disease detection using convolution neural network. Procedia Computer Science. 167: 293-301.

Alsmadi, M.K. (2020). Content-based image retrieval using colour, shape and texture descriptors and features. Arabian Journal for Science and Engineering. 45(4): 3317-3330.

Cambuí, B.G., Mantovani, R.G. and Cerri, R. (2021). Exploring autoencoders for feature extraction in multi-target classification. Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN). IEEE. 1-8.

Chen, Y., Ling, M., Liu, Y., Chen, X., Li, Y. and Tong, B. (2024). Enhancing MRI image retrieval using autoencoder-based deep learning: A solution for efficient clinical and teaching applications. Journal of Radiation Research and Applied Sciences. 17(3): 100932.

El-Habil, A.M. (2012). An application on multinomial logistic regression model. Pakistan Journal of Statistics and Operation Research. 8(2): 271-291.

Hayit, T., Endes, A. and Hayit, F. (2024). KNN-based approach for the classification of fusarium wilt disease in chickpea based on colour and texture features. European Journal of Plant Pathology. 168(4): 665-681.

Jeong, H.Y. and Na, I.S. (2024). Efficient faba bean leaf disease identification through smart detection using deep convolutional neural networks. Legume Research. 47(8): 1404-1411. doi: 10.18805/LRF-798.

Jianing, G., Zhiming, X., Rasheed, A., Tiancong, W., Zhao, Q.I.A.N., Zhang, Z.H.U.O., Zhao, Z.H.U.O., Gardiner, J.J., Ahmad, I., Xiaoxue, W. and Wei, J.I.A.N. (2022). CRISPR/Cas9 applications for improvement of soybeans, current scenarios and future perspectives. Notulae Botanicae Horti Agrobotanici Cluj-Napoca. 50(2): 12678-12678.

Jiawei, H. and Micheline, K. (2006). Data mining: Concepts and techniques. Morgan kaufmann.

Kalpana, M., Karthiba, L., Senguttuvan, K. and Parimalarangan, R. (2023). Diagnosis of major foliar diseases in black gram (*vigna mungo* l.) using convolution neural network (CNN). Legume Research. 46(7): 940-945. doi: 10.18805/LR-5083.

Kappali, H.R., Sadyojatha, K.M. and Prashanthi, S.K., (2024). Computer vision and machine learning in paddy diseases identification and classification: A review. Indian Journal of Agricultural Research. 58(2): 183-187. doi: 10.18805/IJARe.A-6061.

Kotwal, J., Kashyap, R. and Pathan, M.S. (2024). An India soyabean dataset for identification and classification of diseases using computer-vision algorithms. Data in Brief. 53: 110216.

Li, P., Pei, Y. and Li, J. (2023). A comprehensive survey on design and application of autoencoder in deep learning. Applied Soft Computing. 138: 110176.

Maggipinto, M., Masiero, C., Beghi, A. and Susto, G.A. (2018). A convolutional autoencoder approach for feature extraction in virtual metrology. Procedia Manufacturing. 17: 126-133.

Miao, J., Wang, B., Wu, X., Zhang, L., Hu, B. and Zhang, J.Q. (2019). Deep feature extraction based on Siamese network and auto-encoder for hyperspectral image classification. Proceedings of 2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE. pp 397-400.

Na, M.H. and Na, I.S. (2024). Detection and classification of wilting in soybean crop using cutting-edge deep learning techniques. Legume Research. 47(10): 1723-1729. doi: 10.18805/LRF-797.

Nandhini, N. and Bhavani, R. (2020). Feature extraction for diseased leaf image classification using machine learning. Proceedings of 2020 International Conference on Computer Communication and Informatics (ICCCI). IEEE. pp 1-4.

Nandhini, N. and Srisathya, K.B. (2021). May. Identification of plant leaf diseases using adaptive neuro fuzzy classification. Journal of Physics: Conference Series. 1916(1): 012008.

Polic, M., Krajacic, I., Lepora, N. and Orsag, M. (2019). Convolutional autoencoder for feature extraction in tactile sensing. IEEE Robotics and Automation Letters. 4(4): 3671-3678.

Prabaharan, T., Periasamy, P. and Mugendiran, V. (2020). Studies on application of image processing in various fields: An overview. IOP Conference Series: Materials Science and Engineering. 961: 1-13.

Ranjan, M., Weginwar, M.R., Joshi, N. and Ingole, A.B. (2015). Detection and classification of leaf disease using artificial neural network. International Journal of Technical Research and Applications. 3(3): 331-333.

Saradhambal, G., Dhivya, R., Latha, S. and Rajesh, R. (2018). Plant disease detection and its solution using image classification. International Journal of Pure and Applied Mathematics. 119(14): 879-884.

Shukla, P.K. and Sathiya, S., 2022, June. Early detection of potato leaf diseases using convolutional neural network with web application. Proceedings of 2022 IEEE World Conference on Applied Intelligence and Computing (AIC). IEEE. 277-282.

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): 19141.

Xing, C., Ma, L. and Yang, X. (2016). Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. Journal of Sensors. 2016(1): 3632943.

Yao, J., Tran, S.N., Sawyer, S. and Garg, S. (2023). Machine learning for leaf disease classification: data. techniques and applications. Artificial Intelligence Review. 56(3): 3571-3616.