Title Page 2

Enhancing accuracy in tea leaf disease detection by a comparative analysis of convolutional neural networks (CNNs) and random forest

V.Mahitha[1],Dr.K.Anbazhagan[2]

V.Mahitha[1]

Research Scholar,

Department of Computer Science and Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical And Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India. Pincode:602105.

vardhireddymahitha1440.sse@saveetha.com

Dr.K.Anbazhagan[2]

Research Guide, Corresponding Author,

Department of Computer Science and Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical And Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105.

Keywords:machine learning, novel convolutional neural networks (CNNs), novel random forest, sustainable agriculture,resilient agriculture,disease identification

ABSTRACT

Aim: We seek to improve the accuracy of tea leaf disease detection by comparing random forest and convolutional neural network models, looking at their performance metrics, and fine-tuning the chosen model for real-world use.
Materials and Methods: Using artificial neural networks (N = 20) and convolutional neural networks (N = 20), the tea leaf disease was identified**.** Each of the two groups that comprise the original data has a sample size of twenty. The G-power pretest of 80%, threshold of 0.05%, and confidence interval (CI) of 95% are used in this study's final sample size computation. Because of its great accuracy, the Convolutional Neural Network Algorithm, a supervised machine learning and deep learning recognition technique, is essentially required for classification and recognition. Random Forest is a powerful ensemble learning method for tea leaf disease detection. **Results:** The precision of tea leaf disease identification using CNNs is 94%, and that of random forest is 78.95%. There's a huge distinction between convolutional neural network techniques (CNNs) and random forest (0.003 ($p<0.001$)). **Conclusion:** Convolutional neural networks (CNNs) seem more accurate than random forest algorithms in tea leaf disease detection.

Keywords: machine learning, novel convolutional neural networks (CNNs), novel random forest, sustainable agriculture,resilient agriculture,disease identification

INTRODUCTION:

Tea is one of the most popular drinks in the world and has significant cultural and commercial value. However, a number of diseases that affect tea plants pose severe challenges to the tea industry, resulting in significant losses in both the amount and quality of tea (Soeb et al. 2023).Using state-of-the-art machine learning techniques to detect tea leaf disease not only increases agricultural productivity but also supports sustainable agriculture practices and contributing to the resilient agriculture.. For these diseases to be effectively managed and preventive measures implemented, a prompt and correct diagnosis is essential (Pandian et al. 2023).In addition to maintaining tea output, the timely detection of diseases in tea leaves is essential to protecting the livelihoods of millions of people working in the tea industry (Zhao et al. 2022). Tea crops are susceptible to several diseases, which emphasises the need for sophisticated diagnostic methods and ongoing research endeavours to lessen the consequences of these illnesses (Heng, Yu, and Zhang 2024)

Approximately 346 papers have been uploaded to IEEE resources, and over the last four years, 78 papers have been uploaded to Google Updated Scholar. (Chaliha et al. 2022) with an approximate success rate of roughly 90% (Huang et al. 2022) have presented a new model for tea leaf disease detection called a novel convolutional neural network model. employed improvised models and random forests to diagnose diseases, yielding more accurate results than current methods. (Pan et al. (2024) employed the random forest technique to identify hidden patterns in the dataset in order to more effectively classify the data. The highest accuracy of over 89% was attained. (Sattler et al. 2020) a previously owned fuzzy technique to diagnose problems with tea leaves. It combines support vector machines with innovative convolutional neural network methods, with a maximum accuracy of about 94%.

In terms of tea leaf disease identification, all of the long-running deep learning and machine learning models perform less accurately. In light of this, the current paper's goal is to identify activities using random forests and the Novel Convolutional Neural Network Algorithm that produce noticeably improved and more accurate results. By making changes to the models and utilising a larger dataset with more parameters and diverse results, these methods considerably increase the recognition of tea leaf disorders in comparison to previous models (Yang et al. 2023).The objective is to improve the random forest's accuracy rate in the diagnosis of tea leaf sickness by employing a new and improved convolutional neural network technique.

MATERIALS AND METHODS

The proposed project is conducted in the Computer Vision Department of SIMATS, Chennai, which is part of the Saveetha School of Engineering. The Convolutional Neural Network (CNN) and the Support Vector Machines (SVMs) are the two groups involved in this project. The sample size was computed using the Sample Size Calculator (clincalc.com) with a level of 0.07, a G capacity of 365%, and a certainty range of 99.0% based on the findings of previous studies.

The dataset being used at the moment is called train data, and it was obtained from Kaggle.com. There are 885 files in the database, organised into 8 classes. 177 files were used in the validation. They include information about tea leaves that are unhealthy, such as those with "algal," "anthrone," "bird eye," "brown blight," "grey light," "red spot," "white spot," and "healthy" tea leaves.

The Windows 11 OS was used as the evaluation platform. Six gigabytes of RAM and a Ryzen i5 processor comprised the hardware setup. 64-bit system sorting was employed. The Python programming language was used to implement the code. In this case, the algae, anthrone, and bird eye are the dependent variables, while a healthy leaf is the independent variable. To increase accuracy, the analysis is conducted using both independent and dependent variables. In the dataset, 80% was used for training and 20% for testing.

Convolutional Neural Networks (CNNs)

CNNs, or artificial neural networks, are a particular kind of network designed to process structured grid data, primarily for computer vision applications. The convolutional, pooling, fully connected, and activation layers on CNNs enable them to extract features from input images in a hierarchical and adaptive manner. Fully connected layers generate final outputs, pooling layers reduce dimensionality, and convolutional layers use filters to capture spatial hierarchies. The use of non-linear activation functions complicates the learning process. CNNs are extremely effective at tasks like object identification, facial recognition, and image classification because they were trained with labelled datasets and the backpropagation method. CNNs are a key component of artificial intelligence and machine learning, particularly in the area of visual data processing, since they make it easier to customise them for a given task by using transfer learning with pre-trained models.

algorithm:

Step 1. Gathering the dataset (of tea leaves containing diseased tea leaves and a set of healthy tea leaves)
Step 2: Preparing the data
Step 3. Labelling the item
Step 4: Construct the architecture for CNN. Convolutional layers, pooling layers for feature extraction, and fully connected layers for classification are the common building blocks of a typical architecture. Add suitable output layers and activation functions, such as ReLU.
Step 5: Model Assembly
Step 6. Divide the data set into sets for validation and training.
Step 7: Evaluate how well the model performs on a test set that was not used for training to see how well it can adapt to new data.

Step 8. Once satisfied with the model's performance, deploy it for real-world use

Random Forest

The process of tea leaf disease detection using the novel Random Forest algorithm commences with the initialization of an empty list called Random_Forest. Decision trees are then methodically constructed through iterations. Each time, a random subset from the training dataset is used to build a decision tree. The resulting tree Next, the tree that is produced is added to the Random_Forest list; a second random subset is produced, and forecasts are derived separately from every decision tree in the Random Forest. A principal majority voting system totaling the predictions from each individual tree determines the final prediction. When stopping conditions are satisfied (such as reaching a maximum depth or having a minimum number of samples per leaf), leaf nodes are generated as part of the decision tree construction phase. Furthermore, the best splits are used to construct decision nodes recursively. Leaf nodes indicate the dataset's majority class of samples, whereas decision nodes hold vital information about specific features, splits, and subtrees. This all-encompassing algorithm aims to improve the precision of Using the cooperative decision-making abilities of multiple trees within the Random Forest framework for tea leaf disease identification

Pseudocode

Step 1: Finding the dataset with the images of tea leaves and the associated disease labels.

Step 2: The desired number of trees in the random forest.

Step 2: Determine how many trees you want in the random forest.

Step 3: A grouping of decision trees. Initialise an empty Random_Forest list.

Step 4: Assemble the training dataset into a random subset (subset).

Step 5: Using the random subset, construct a decision tree (tree).

Step 6: Include the decision tree in the list called Random Forest.

Step 7: Assemble a subset at random (subset) in order to make predictions.

Step 8: The Random Forest's trees are as follows:

Step 9: Use the decision tree to make a prediction.

Step 10: Gather the forecasts.

Step 11: Return the final prediction, which is the majority vote of all the predictions.

Step 12: Assign the leaf node to the sample class that makes up the majority of the dataset.

Step 13: Based on the best split, create a decision node.

Statistical Analysis

The IBM SPSS 26.0.1 program is used to perform the statistical analysis for both the compared and proposed models. This dataset contains independent photos of healthy tea leaf disease. Algae, anthrone, and bird eye are the dependent variables. An independent sample T test analysis was conducted. On both models, this analysis is conducted. The standard deviation and standard error are noted following the analysis of the mean accuracy.

RESULTS

Table 1 shows the description of analytics for precision for both algorithms. Convolutional neural network and random forest. Table 2 shows a collective analysis that gives a precision average of 94% for the convolutional neural network technique, which seems more precise in comparison to the random forest approach, which has only 78.95%. Standard deviation and mean errors are calculated (Standard error mean for the convolutional neural Algorithm is 0.25 and random forest Algorithm is 0.16) Table 3 indicates the significance of testing data findings at 0.003 (less than 0.005). The bar chart in Fig. 1 shows the mean accuracy between the convolutional neural network algorithm and the random forest algorithm. From the results, it is clearly evident that the novel convolutional neural network algorithm is performing better when compared to the random forest algorithm.

DISCUSSION

In the current study, we observed that the deep learning, machine learning, and convolutional neural network algorithms appear to have a higher success rate than the random forest ($p = 0.001$, stratified random analysis). The enhanced precision of this CNN (mean precision = 86.4100) is greater than that of a random forest (mean precision =70.4320).

Numerous techniques, including random forests and innovative convolutional neural networks, are used in the investigation (Xiong et al. 2022). When comparing the novel CNNT to the random forest, the novel CNNT performs better than the random forest (Xu et al. 2022). The accuracy of the Novel Convolutional Neural Network Algorithm is 94%, while that of Random Forest is 78.95%. (Panyatip et al. 2022). Additionally, a comparison between the random forest technique and the novel CNN algorithm was conducted. The findings indicate that the novel convolutional neural network algorithm outperforms the random forest algorithm. It has also produced outcomes that are comparable to what we ultimately obtained (Zhou and Chen 2018). Additionally, tea leaf infection identification was carried out using supervised pattern recognition and advanced analytics algorithms. The findings indicate that novel convolutional neural network algorithms outperform all other algorithms in terms of accuracy and performance (Nakamatsu et al. 2020). Furthermore, we can draw the conclusion that the novel convolutional neural network method performs better than the random forest technique and appears to be more exact based on the talks and observations above.

Convolutional neural networks have certain limitations when it comes to the detection of tea disease. They are made up of large clusters of CNN classifiers and require longer execution times than other machine learning and advanced analytics algorithms. This model will be improved in later work to have more features, operate faster, and yield the most accurate results. This might have a better future because there are more actions every day.

CONCLUSION

Two distinct algorithms—the random forest algorithm and the convolutional neural network algorithm—are used in this paper to identify tea leaf disease. Compared to the random forest algorithm, the convolutional neural network algorithm performs better statistically and with a higher precision level.

DECLARATIONS

Conflicts of interests

Conflicts of interest are not present in the manuscript.

Authors Contribution

In addition to writing the manuscript, author V. Mahitha was involved in data collection, data analysis, and a literature review. Data validation, verification, and manuscript review are all done by the author, K. Anbazhagan.

Acknowledgements

The authors would like to express their gratitude to Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing the infrastructure required to successfully complete this research.

Funding

We thank the following organisations for providing financial support that enabled us to complete the study:.

1. Byju's Learning Academy
2. Saveetha University

3. Saveetha Institute of Medical and Technical Sciences

4. Saveetha School of Engineering

REFERENCES

1.Chaliha, Chayanika, Richa Srivastava, Eeshan Kalita, Lingaraj Sahoo, and Praveen Kumar Verma. 2022. "Rapid and Precise Detection of Cryptic Tea Pathogen Exobasidium Vexans: RealAmp Validation of LAMP Approach." *World Journal of Microbiology & Biotechnology* 39 (2): 52.

2.Heng, Qidong, Sibo Yu, and Yandong Zhang. 2024. "A New AI-Based Approach for Automatic Identification of Tea Leaf Disease Using Deep Neural Network Based on Hybrid Pooling." *Heliyon* 10 (5): e26465.

3.Huang, Minglei, Qiong Wu, Xing Wang, Thangaraj Kuberan, Fuxing Shu, Gregory Joseph Duns, and Jishuang Chen. 2022. "First Report of Blight Caused by AG4-HGI on in Guizhou, China." *Plant Disease*, August. https://doi.org/10.1094/PDIS-05-22-1255-PDN.

4.Nakamatsu, Kazumi, Roumen Kountchev, Ari Aharari, Nashwa El-Bendary, and Bin Hu. 2020. *New Developments of IT, IoT and ICT Applied to Agriculture: Proceedings of ICAIT 2019*. Springer Nature.

5.Pandian, J. Arun, Sam Nirmala Nisha, K. Kanchanadevi, Abhay K. Pandey, and Samira Kabir Rima. 2023. "Grey Blight Disease Detection on Tea Leaves Using Improved Deep Convolutional Neural Network." *Computational Intelligence and Neuroscience* 2023 (January): 7876302.

6.Pan, Xiangming, Yixuan Zhou, Dongxue Liu, Yiyi Wang, Yuting Sheng, and Hong-Xia Zhang. 2024. "Leaf Spot Caused by on in China." *Plant Disease*, February. https://doi.org/10.1094/PDIS-11-23-2342-PDN.

7.Panyatip, Panyada, Tanit Padumanonda, Chawalit Yongram, Tiantip Kasikorn, Bunleu Sungthong, and Ploenthip Puthongking. 2022. "Impact of Tea Processing on Tryptophan, Melatonin, Phenolic and Flavonoid Contents in Mulberry (L.) Leaves: Quantitative Analysis by LC-MS/MS." *Molecules* 27 (15). https://doi.org/10.3390/molecules27154979.

8.Sattler, Kai-Uwe, Duy Cuong Nguyen, Ngoc Pi Vu, Banh Tien Long, and Horst Puta. 2020. *Advances in Engineering Research and Application: Proceedings of the International Conference on Engineering Research and Applications, ICERA 2019*.

9.Soeb, Md Janibul Alam, Md Fahad Jubayer, Tahmina Akanjee Tarin, Muhammad Rashed Al Mamun, Fahim Mahafuz Ruhad, Aney Parven, Nabisab Mujawar Mubarak, Soni Lanka Karri, and Islam Md Meftaul. 2023. "Tea Leaf Disease Detection and Identification Based on YOLOv7 (YOLO-T)." *Scientific Reports* 13 (1): 6078.

10.Xiong, Haoyang, Meifeng Chi, Hui Lyu, Chengjian Xie, Ling He, Xing Yang, Wei Wei, et al. 2022. "First Report of Phyllosticta Capitalensis Causing Leaf Spot of Mahonia Fortunei in China." *Plant Disease*, November. https://doi.org/10.1094/PDIS-05-22-1054-PDN.

11.Xu, Jia, Yang Wei, Fanglan Li, Xinchu Weng, and Xinlin Wei. 2022. "Regulation of Fungal Community and the Quality Formation and Safety Control of Pu-Erh Tea." *Comprehensive Reviews in Food Science and Food Safety* 21 (6): 4546–72.

12.Yang, Yishuai, Xueyu Yang, Yudan Zhang, Zuohua Ren, Jie Zhong, Qiulong Hu, and Lin

Tan. 2023. “First Report of Causing Root Rot of Tea Plants () in China.” *Plant Disease*, September. https://doi.org/10.1094/PDIS-06-23-1172-PDN.
13. Zhao, Xiaohu, Jingcheng Zhang, Ailun Tang, Yifan Yu, Lijie Yan, Dongmei Chen, and Lin Yuan. 2022. “The Stress Detection and Segmentation Strategy in Tea Plant at Canopy Level.” *Frontiers in Plant Science* 13 (July): 949054.
14. Zhou, Jianlong, and Fang Chen. 2018. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Springer.

TABLES AND FIGURES

Table 1. Accuracy values of ann and random forest k

S. No	CNN	Random Forest
1	94.00	78.95
2	93.00	77.00
3	91.60	75.88
4	92.00	74.00
5	90.80	7322
6	89.80	72.50
7	88.70	71.66
`8	87.00	70.00
9	92.90	69.11
10	45.00	42.00

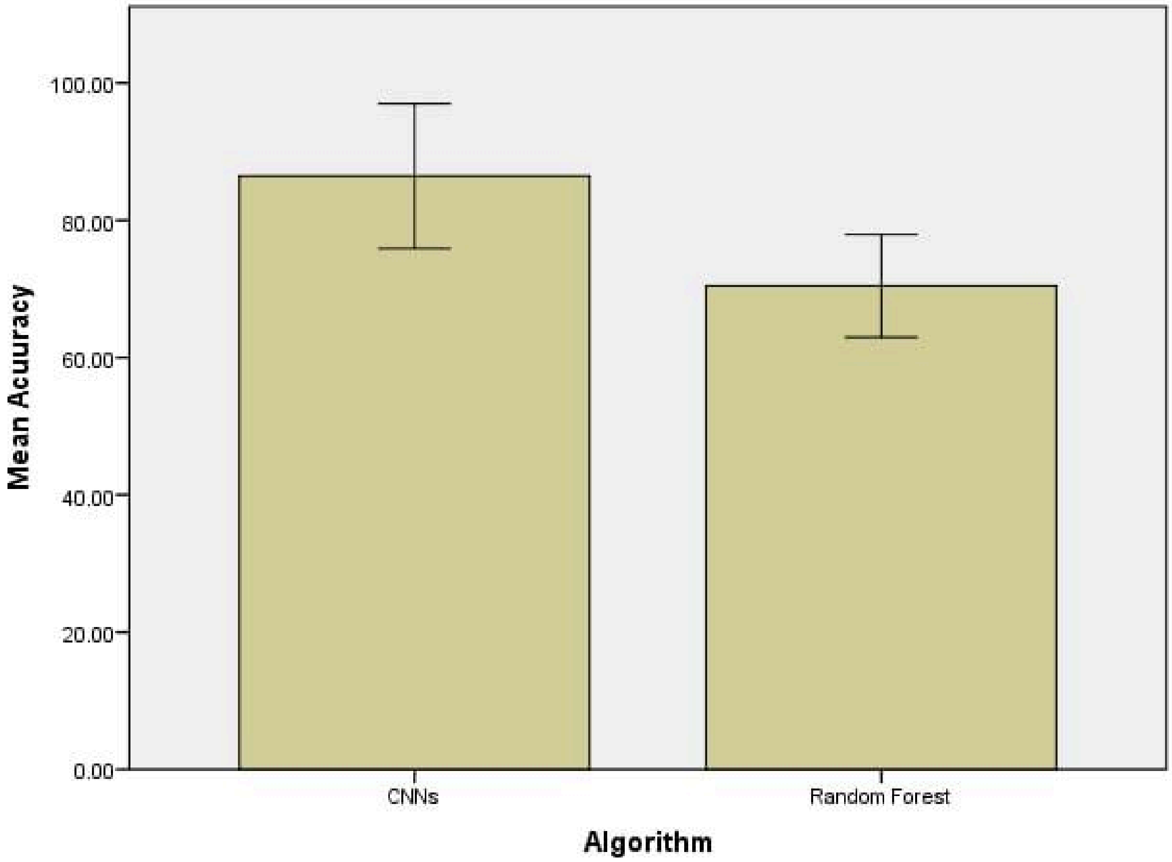
Table 2. Group Statistics Results (The mean of the neural network algorithm is 86.4100, which is more compared to the random forest algorithm, 70.4320, where the 70.4320 where

Average variances for CNN technique is 0.25 and random forest   algorithms 0.16).

	GROUPS	N	Mean	Std. Deviation	Std. Error Mean
ACCURACY	Convolutional neural networks	10	86.4100	14.71873	4.65447
	Random forest algorithm	10	70.4320	10.44550	3.30316

Table 3: Individual Data Analysis to Determine Importance and Variance Average. A P rate is 0.003 (lower than 0.005) was determined to be statistical analysis and 95% confidence interval was considered.

		Levene's test for Equality of Variances		**T-test for equality of means**						
		F	**sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence interval of the difference**	
									Lower	**Upper**
Accuracy	**Equal variances assumed**	.241	.629	2.800	18	.012	15.97800	5.70745	3.98710	27.96890
	Equal variances not assumed			2.800	16.231	.013	15.97800	5.70745	3.89275	28.06325



Error Bars: 95% CI
Error Bars: +/-2 SD