Cacao Plant Disease Detection and Classification

Emmanuel Tovurawa
u20769271@tuks.co.za

Department of Computer Science, University of Pretoria, South Africa

Blessing Dlamini
Department of Computer Science, University of Pretoria, South Africa

Stephen E. Moore
Department of Mathematics, University of Cape Coast

Vukosi Marivate
Department of Computer Science, University of Pretoria, South Africa

Abiodun Modupe
Department of Computer Science, University of Pretoria, South Africa

---

Research Article

Keywords: Machine learning, Deep learning, computer vision, convolution neural network

Posted Date: January 7th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5763786/v1

Additional Declarations: The authors declare no competing interests.

Cacao Plant Disease Detection and Classification

Emmanuel Tovurawa[1], Blessing Dlamini[1], Stephen E. Moore[2,3], Vukosi Marivate[1], and Abiodun Modupe[1]

[1] Department of Computer Science, University of Pretoria,
[2] Department of Mathematics, University of Cape Coast, Ghana
[3] Center for Artificial Intelligence Research, Ghana

Abstract. Food security is a vital aspect of the United Nations' Sustainable Development Goals (SDGs) which aims to promote sustainable farming in the world. Farming-driven economies such as Ghana are faced with challenges due to plant diseases. Cacao, a vital crop in Ghana is severely impacted with diseases which affect its yield and decrease exports revenue through reduced exports. Leveraging deep learning techniques offers an effective solution for early detection of diseases in cacao plants. This study adopts a comprehensive approach, starting with an Exploratory Data Analysis (EDA) of the dataset containing images of both healthy and diseased cacao plants from Ghanaian farms. Using exploratory data analysis (EDA), we can identify patterns and understand the characteristics of the dataset, laying a solid foundation for developing robust machine learning models tailored to the specific challenges faced by Ghanaian cacao farmers. Our approach involves developing and evaluating deep learning models to detect and classify cacao plant diseases. These models are designed with the Predictability, Compatibility, and Stability (PCS) framework in mind, ensuring reliability and effectiveness in disease detection. The custom convolution neural network (CNN) model outperformed other models considered in experimental analysis. This study aims to revolutionize cacao farming through precise, stable, and ethical deep learning solutions, ultimately enhancing crop resilience, productivity, and the livelihood of Ghanaian farmers.

Keywords: Machine learning · Deep learning· computer vision· convolution neural network· model · visualization.

1 Introduction

Cocoa, a key ingredient in chocolate, cosmetics, and health products, is one of Ghana's most important agricultural exports, making the country one of the largest exporter of cocoa globally [3]. However, the declining trend in cocoa production is attributable to various diseases affecting cacao plants, highlighting the urgent necessity for effective tools to detect and control plant diseases early [2]. Currently, farmers lack sufficient and affordable instruments for timely disease detection and classification, impacting their yield and income [11].

This project focuses on the development of a deep learning model for the detection and classification of cacao plant leaf diseases, utilizing a dataset that includes images of both healthy and diseased plant leaves from Ghanaian farms and open-source repositories. The initial phase involves Exploratory Data Analysis (EDA) to uncover insights and patterns within the dataset, which is crucial for building accurate and reliable models. A comparative analysis of deep learning techniques was performed by considering custom Convolution Neural Network (CNN), ResNet50, and EfficientNetB0 models. The models are developed and evaluated using the Predictability, Compatibility, and Stability (PCS) framework to ensure they are of high quality. Due to its outstanding performance, the custom CNN model is selected and implemented in this project.

The primary problem addressed by this article is the need for a reliable and cost-effective solution for early detection and classification of cacao plant diseases. By leveraging deep learning techniques, this project aims to develop a model capable of accurately identifying diseased cacao plant leaves using image data. This model can be deployed in a user-friendly application that provides real-time diagnostic feedback to farmers, enabling them to manage plant health more effectively and improve their crop yields.

The main research question is: How effective are convolution neural networks (CNNs) in detecting diseases in cacao plants? To address this, we break it down into three sub-questions:

1. How can you minimize false positives and achieve high accuracy?
2. What role does hyper-parameter fine-tuning play in building a robust model?
3. How does the quality and diversity of the training dataset impact the performance and generalization ability of the disease detection model?

These questions will guide the development of an accurate model, contributing to the advancement of cacao plant disease detection and classification, ultimately improving food security and the global supply.

2 Related Work and Literature Survey

2.1 Introduction to Plant Disease Detection

Plant diseases threaten the world's food security and have a significant detrimental impact on agricultural productivity. Early disease diagnosis and treatment are critical to preserving sustainable agriculture and increasing crop yields. Traditional disease detection methods are expert examination-based, expensive [14] and time-consuming, which large-scale farmers cannot afford [8], [11]. Although the manual method of identifying and tracking diseases in cacao leaves is the standard approach, it necessitates a significant amount of labor, knowledge of plant diseases, and lengthy processing times [11], [15]. As a result, modern approaches that leverage machine learning (ML) and deep learning (DL) techniques have gained popularity and offer automated, accurate, and timely plant disease detection solutions. An interesting discovery that has the potential to drastically change the cocoa farming environment is the incorporation of AI through CNNs to detect these agricultural challenges in Ghana's cocoa estates [2]. This technology promises to manage crop health in a more accessible, precise, and effective manner, which will not only increase productivity but also assure the sustainability of cocoa cultivation in the face of changing difficulties.

2.2 Body

Machine learning techniques are extensively employed in agriculture to perform tasks such as soil analysis, crop yield prediction, and disease identification [7]. With the development of technology came techniques for the quicker and simpler identification of plant diseases, such as remote sensing, image segmentation, and soft computing [11]. Machine learning approaches such as Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVM), and k-Nearest Neighbours (k-NN) have shown promise in the classification and detection of plant diseases [4]. While plant disease symptoms can be effectively classified using machine learning (ML), the techniques requires several labour intensive feature extraction and pre-processing steps that are less adaptable when utilized with different datasets [4]. Deep learning's ground breaking work, particularly with regard to its use of convolution neural networks (CNNs), which have revolutionized the field of computer vision, has tremendously aided in the diagnosis of plant diseases. Natural language processing, speech processing, and picture and video processing have all made extensive use of deep learning [13]. Convolutional neural networks is a subset of artificial neural networks (ANN) currently widely utilized for object recognition, segmentation, and picture classification [7]. Numerous industries, including health, agriculture, the web, mail services, etc., have extensively adopted CNN. A neural network can identify patterns, categorize data, and predict future events by learning from historical data [7]. CNNs are ideal for image-based disease detection applications because they automatically extract features from raw data, eliminating the need for manual feature engineering [1], [3]. Deep learning models have shown to be more accurate and resilient in a range of environmental conditions and plant species when it comes to diagnosing plant diseases from pictures.

The field of work on CNN-based disease detection in plants is very new, yet it is expanding significantly [2], [9]. CNNs, which represent deep learning paradigms, have proven crucial in helping farmers overcome challenges related to crop cultivation. By examining images of plant leaves, CNNs can accurately detect the signs of plant diseases. This approach makes use of CNNs' ability to analyse large volumes of visual data and spot complex patterns, which makes them ideal for spotting subtle medical signs that could be challenging for human professionals to detect. Plant diseases have recently been categorized using a number of CNN architectures, such as AlexNet [12], VGGNet, ResNet [5], [8], EfficientBO, LeNet-5, VGG-16 and Custom CNN [9], [10]. An example study that used CNNs to classify cocoa pods demonstrated the effectiveness of the EfficientNetB0 model and proposed it as the best model for that particular task [9].Although the depth and complexity of these designs vary, fully connected layers are used for classification, pooling layers are used to reduce dimensionality, and convolution layers are used to extract features. The efficacy of these models in other plant disease scenarios suggests that a similar adaption of these algorithms could be advantageous for the identification of cacao plant diseases.

Developing a CNN-based model for plant disease diagnosis involves several important steps. First, exploratory data analysis (EDA) is performed in order to understand the dataset [11], [13]. Pre-processing tasks including noise reduction and normalization are carried out, as well as an analysis of the distribution of images across different disease categories and a visual inspection of disease symptoms [3]. A variety of CNN models are trained and evaluated using EDA in order to determine the best architecture. To enhance the model's performance and generalizability, methods such as hyper-parameter tuning, data augmentation, and cross-validation are applied.

One of the main obstacles to using CNNs for plant disease identification is the need for vast and diverse information [3]. Acquiring good-quality images of plants that are disease-ridden and healthy is a challenge. In order to expand the variety and size of datasets on purpose, data augmentation techniques have been applied to improve the accuracy and robustness of the models.

The efficacy of artificial intelligence (AI) models can be affected by variable field conditions, such as weather and lighting. The influence of these environmental factors on disease detection based on images required models that are able to adapt to these variations. Advances in transfer learning offer practical means of adapting previously trained models to fit novel environments, enhancing their applicability in numerous agricultural scenarios [3].

For AI technology to be extensively employed in the agriculture sector, it must be simple to use and accessible to farmers [5]. AI technology can greatly improve crop health and output by offering prompt and accurate plant disease identification. In addition to increasing productivity, this technology integration minimizes crop losses and cuts down on the need for chemical treatments, which supports the sustainability of cacao farming as well. There are significant wider ramifications for this technology. As mentioned in [2], comprehensive monitoring systems can be created by combining ground-level observations with data from remote sensing, including satellite images. These systems have the potential to offer insightful information about crop health, facilitating more focused and efficient interventions.

The application of CNN-based deep learning algorithms to the diagnosis of plant diseases in cacao shows great potential for improving farming methods [9]. These models can support sustainable farming methods, lessen crop losses, and lessen the need for chemical pesticides by offering precise, fast, and automated disease identification. To further improve these technologies and extend their applicability to a wider range of crops, including cacao plant diseases, ongoing research and development in this field is essential. The revolutionary potential of CNNs in plant disease detection is emphasized in this literature review, which also emphasizes the significance of data quality, model optimization, and accessibility in order to optimize the advantages of AI in agriculture. We seek to add on the literature with our extensive model.

Our study introduces a novel cacao disease detection and classification model. Through this comprehensive literature review, we identified gaps in current research and developed a light weight custom CNN model that can be used by farmers for cacao plant leaf disease and classification. Our contribution lies in advancing existing methodologies and providing new insights into disease detection and classification in Cacao plants, thereby enriching the current body of research in this domain.

3 Methodology

Goal setting forms a key element in project scoping which is the initial stage of a project. Goal and needs analysis were done as a way of understanding the problem statement of the project. The main objective of the project is to build a cacao plant leaf detection and classification model to identify cacao leaf diseases. The next step of the project was to decide on the relevant methodology to employ in our study. Deep learning models such as Convolution Neural Networks (CNNs) are designed to process and analyze data. There are wide variety of architectures which exist in CNNs therefore it was necessary to do experimental and comparative analysis. As such, this was the methodology approach used. Input data for our project are images of both healthy and diseased cacao plant leaves. These images were obtained from different sources and are representative of all farming regions in Ghana. There are three classes in the dataset namely anthracnose, cocoa swollen shoot virus and healthy cocoa. A sample image from each class is shown below.

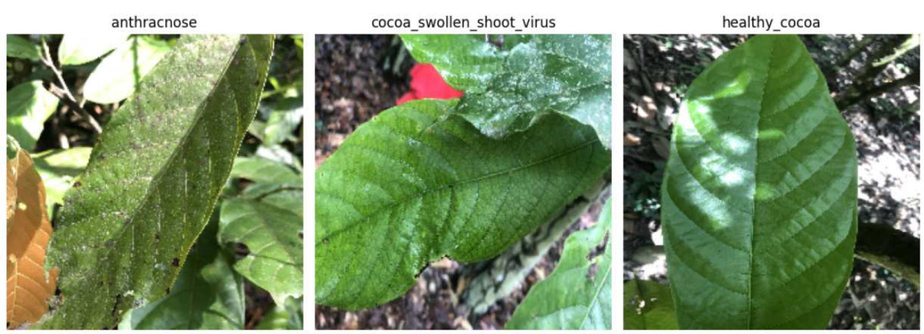


Fig. 1: (left) Image of cacao leave affected by Anthracnose, (middle) image of cacao leaf affected by swollen shoot and (right) healthy cacao leaf

It can be seen from images above that at times an image can have multiple leaves in it and this was important to have in our model as it is reflective of the true reality in the farms. When a farmer takes an image to classify it, there may be other cacao leaves in the background therefore model should be able to handle such cases. Exploratory Data Analysis (EDA) is a process of analysing input data so as to gain understanding which will help inform modelling decisions. This will be discussed in more detail in the next section .

Data pre-processing helps identify outliers and possible issues with the dataset. This involves data cleaning which is necessary to ensure data quality is good so that sound and accurate models can be developed. There are other aspects also considered such as image augmentation, data transformation, data splitting. Image augmentation is the process of making adjustments such as rotation, flipping, cropping to achieve better quality images. Data transformation is the process of making changes to the images for example resizing and normalisation. The image target size defined in the model is 224 pixels so all images were resized to 224pixels. Image normalisation is performed so as to scale pixel values to a standard range (0,1), where zero represents the mean and one is the variance. This helps ensure that all image features contribute equally to the model minimising bias. Data splitting is a fundamental step in model building and this involves creating a subset of the data which will be used for training, validation and testing purposes. Our data is split into 70% and 30% for training and testing purposes. The training dataset is further split into 70% and 30% for training and validation purpose. The training dataset is used for building the model and it is during the phase when the model learns patterns and features are extracted. Validation dataset is used for hyper-parameter fine tuning and monitoring of the model performance ensuring there is no over-fitting. The testing dataset helps provide an evaluation of the final model performance. Testing the model with unseen data ensures that there is no bias in model performance. Splitting of the data helps deal with data leakage and generalization ability of the model.

Model Building Given the abundant CNN architectures, we explored three different ones namely ResNet50, custom CNN and EfficientNetB0. The key objectives in the model development phase are as follows:

- Develop a cacao plant leaf disease detection and classification model which can be used by farmer.
- Minimize false positives and have high prediction ability.

[2] discussed the potency of CNNs in disease detection and further delved into intersection of agriculture and Artificial Intelligence (AI). Convolution neural networks have the ability to capture spatial hierarchies in the data due to the different layers which they have. There are four main layers which are convolution, pooling, fully connected and output layer.

An image is inserted into the neural network and this is of size (224,224,3) where 224 is the image size and three represents RGB color channels. The convolution layer is responsible for extracting features and patterns in the image through kernel filters. Each filter slides over the image and performs element-wise mathematical operations and produces a single output for each region. After that, a feature map is produced and sent to the Rectified Linear Unit (ReLU), where an activation function is used to give the network non-linearity and the ability to recognize intricate patterns. The pooling layer is used for spatial dimensions reduction whilst ensuring important features are retained. Normalising feature maps helps with training stability and convergence. After this, the feature maps are flattened meaning that the multi-dimensional array of features is reshaped to one vector which will go through the fully connected layers. Fully connected layers, also called dense layers help in learning of complex relationships through combining features. Dropout prevents over-fitting therefore its key that the neural network has it. The output layer is responsible for producing the final class prediction.

The models considered in the experimental analysis follow the process described above however they have distinguishing elements and functions. The EfficientNetB0 architecture encompasses compound scaling. This means that network depth, with and resolution are scaled by some parameter. This model is computationally expensive due to it having seven deep layers. ResNet50 architecture is different in that it has fifty dense layers and an extra layer for residual block responsible for addressing disappearing gradients. The final model selected is the customised CNN which performed well compared to the other two. Ensemble learning where all three models were combined was also performed, however, this produced low accuracy. The process flow shown below summarises the end-to-end process followed in this project.

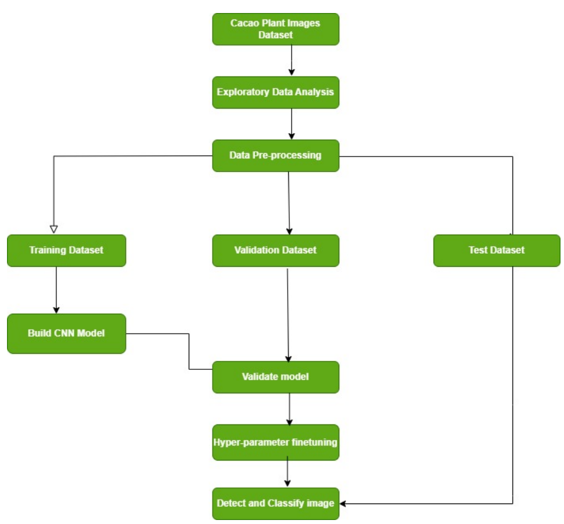


Fig. 2: Process Flow

4 Exploratory Data Analysis (EDA)

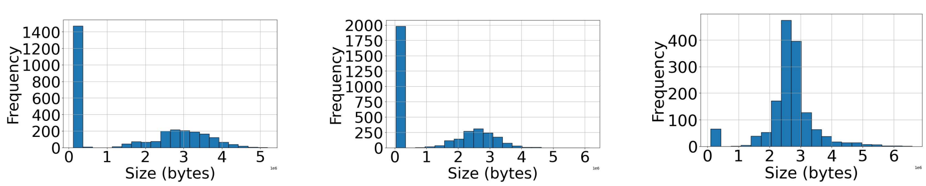
The numerous exploratory data studies of cacao plant leaves data that are pertinent to the development of an effective and precise disease detection model are examined in this subsection.

4.1 Cross Balance Analysis

There are several discrepancies in the data classifications, which could impact the analysis's findings. The cross balance analysis analyzes the distribution of data across different data classes. The results from the analysis indicates that the distribution of images across the cocoa swollen shot virus, the anthracnose and healthy classes is 3422, 2991 and 1495 images respectively. This could skew the model toward majority groups and reduce its accuracy toward minority groups. To increase the model's predicted accuracy, corrective measures like data augmentation, class weighting, or data resampling may be required.

4.2 Distribution of Image Sizes

Image size is an important attribute of the dataset therefore understanding this information is helpful in creating a robust model. It is important that cacao plant leaves sizes are resized to a certain target size. A target image size is specified and all images are transformed to have that size. The bar graphs in Figure 3 below shows the distribution of images sizes for each class.



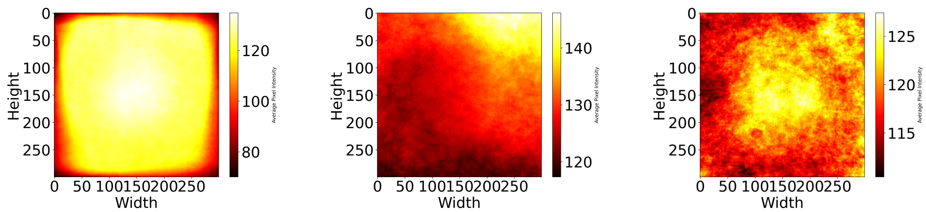
(a) Anthracnose (b) CSSV (c) Healthy

Fig. 3: Image Size Distribution for Classes

Based on the bar graph illustrations there is evidence that the pixel sizes of cacao plant leaves follow the normal distribution. However, there exists some outliers which need to be dealt with.

4.3 Spatial Heatmaps

Spatial heatmaps highlight the regions of an image that contribute most to the detection and classification of image class.



(a) Anthracnose (b) CSSV (c) Healthy

Fig. 4: Spatial Heatmaps for Classes

The heatmaps above show the parts of the image that are indicative of a particular cacao plant leaf disease. The majority, especially the central region of the leaf contributes the most to the detection of anthracnose in cacao plant leaf and it actually represent the majority of the leaf. Although not that significant, the central region of healthy cacao plant leaf is more relevant for the detection/identification of healthy cacao than any other part of the leaf. For cocoa swollen shoot virus, it is the leaf edges which contributes much to the accurate detection of the plant disease. These are the areas of the leaves which the CNN model focuses on much to detect and classify cacao plant leaf diseases.

5 Model Results

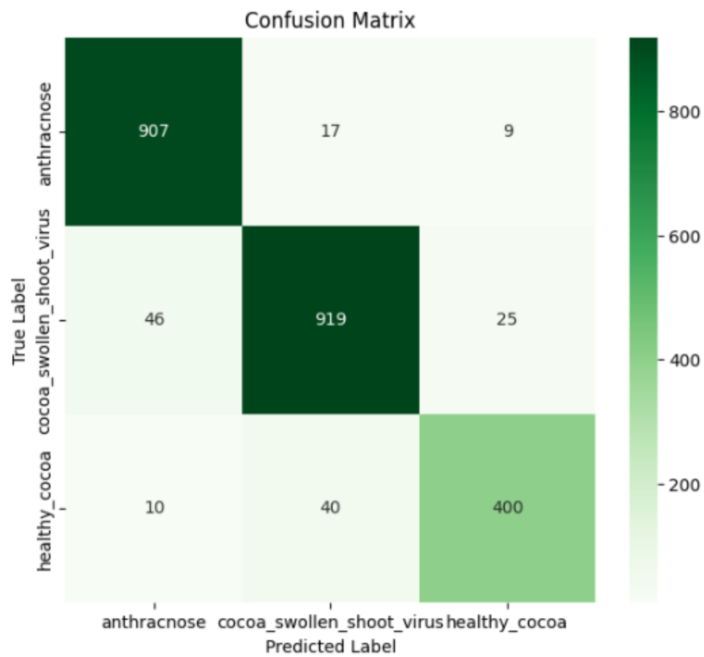
5.1 Model Accuracies

In this section, we discuss model results by looking at two metrics namely accuracy and loss. We evaluated the model's accuracy in terms of three aspects ie training, validation and testing accuracies. The training, validation and testing accuracies after five epochs are 98%, 94% and 94% respectively. The model showed a strong ability to learn and generalize, as indicated by the high accuracy scores. Loss represents a quantitative measure of how well a model's predictions match the true classes in the dataset. It quantifies the difference between predicted classes and the actual ground truth. Lower loss indicates better performance of the model. The final training loss values after five epochs for the model were 0.04, 0.24 and 0.25 for training, validation and testing loss respectively. The decrease in training loss indicates effective learning, while the stabilization of validation loss suggests that the model is not over-fitting and retains its generalization capabilities.

In terms of computability, the model requires approximately 1200 seconds per epoch for training and 250-260 seconds for evaluation, manageable within typical deep learning resource constraints. The model is feasible for deployment with appropriate hardware. Stability is evidenced by consistent accuracy and loss values across epochs, with validation and test accuracies remaining close (93.1% to 94.4%) and validation loss stabilizing around 0.21 to 0.24, indicating robustness. The model's stable performance across different training stages confirms its reliability for practical deployment in cacao leaf disease classification.

5.2 Confusion Matrix

To further assess the model performance, a confusion matrix is used to uncover false negatives and false positives.



	precision	recall	f1-score	support
anthracnose	0.94	0.97	0.96	933
cocoa_swollen_shoot_virus	0.94	0.93	0.93	990
healthy_cocoa	0.92	0.89	0.90	450
accuracy			0.94	2373
macro avg	0.94	0.93	0.93	2373
weighted avg	0.94	0.94	0.94	2373

Fig. 5: Confusion Matrix

The model performs exceptionally well for anthracnose class with high precision (0.94) and recall (0.97), resulting in a high F1-score (0.96). This indicates the model is very good at identifying anthracnose with few false positives and false negatives. The model also performs very well for the cocoa swollen shoot virus class, with precision and recall both at 0.94 and 0.93 respectively, resulting in a balanced F1-score of 0.93. Although slightly lower than the other two classes, the performance for the healthy class is still strong with precision at 0.92 and recall at 0.89, yielding an F1-score of 0.90. This class has the lowest recall, indicating some healthy cacao instances are being misclassified. The model achieves an overall accuracy of 94%, indicating high performance across all classes. The macro averages for precision, recall, and F1-score are 0.94, 0.93, and 0.93, respectively. This suggests consistent performance across all classes, regardless of their individual support counts. The weighted averages are also 0.94 for precision, recall, and F1-score, confirming that the model maintains high performance even when considering the class distribution.

In conclusion, the model is highly effective in accurately identifying and classifying cacao plant leaf diseases, demonstrating strong predictability, balanced performance across classes, and overall high reliability.

6 Conclusion

In conclusion, the development of deep learning model for cacao plant leaf disease detection and classification presents a transformative opportunity for the agricultural sector. By providing farmers with a tool that enables real-time analysis and diagnosis through image uploads, we can significantly enhance the precision and timeliness of disease management. This capability hinges on robust data pre-processing, exploratory data analysis, and effective data visualization to ensure model accuracy and reliability. The research questions posed at the onset of the project have been addressed and answered through meticulous and comprehensive model building. The model performance is remarkable and this was as a result of hyper-parameter fine tuning and good pre-processing techniques employed. The cacao plant leaf disease detection and classification project has successfully developed a robust Convolution Neural Network (CNN) model to accurately identify and classify cacao plant leaves diseases. Overall, this project demonstrates the potential of AI and machine learning in revolutionizing agricultural practices, providing farmers with advanced tools to ensure healthy crops and sustainable farming practices.

References

1. Jackulin, C., Murugavalli, S.: A comprehensive review on detection of plant disease using machine learning and deep learning approaches. Measurement: Sensors, vol. 24, pp. 100441. Elsevier (2022).
2. Miracle, A.: Enhancing Cocoa Crop Resilience in Ghana: The Application of convolution Neural Networks for Early Detection of Disease and Pest Infestations. Qeios (2024).
3. Agbeli, P., Kobina, J., Anokye, K., Hayfron-Acquah, J.: APPLICATION OF DIGITAL IMAGE PROCESSING TECHNOLOGY TO DETECT DISEASES IN COCOA PLANTS. PhD thesis (2022). https://doi.org/10.13140/RG.2.2.15337.57447

4. Jeswani, J., Ahmed, A. S., Khan, Z., Fernandes, R.: Leaf Disease Detection Using Deep Learning. In: International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022), pp. 989–994. Atlantis Press (2023).
5. Okokpujie, K., Okokpujie, I. P., Ayomikun, O. I., Orimogunje, A. M., Ogundipe, A. T.: Development of a Web and Mobile Applications-Based Cassava Disease Classification Interface Using convolution Neural Network. Mathematical Modelling of Engineering Problems, vol. 10, no. 1. (2023).
6. Dani, H., Bhople, P., Waghmare, H., Munginwar, K., Patil, A.: Review on Frameworks Used for Deployment of Machine Learning Model. International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 2, pp. 211–215. (2022).
7. Harvyanti, A. F. M., Baihaki, R. I., Ridlo, Z. R., Agustin, I. H., et al.: Application of Convolutional Neural Network for Identifying Cocoa Leaf Disease. In: 1st International Conference on Neural Networks and Machine Learning 2022 (ICONNSMAL 2022), pp. 283–304. Atlantis Press (2023).
8. Ashwinkumar, S., Rajagopal, S., Manimaran, V., Jegajothi, B.: Automated plant leaf disease detection and classification using optimal MobileNet based convolutional neural networks. Materials Today: Proceedings, vol. 51, pp. 480–487. Elsevier (2022).
9. Soh, K. S., Moung, E. G., Danker, K. J. J., Dargham, J. A., Farzamnia, A.: Cocoa Diseases Classification using Deep Learning Algorithm. In: ITM Web of Conferences, vol. 63, pp. 01014. EDP Sciences (2024).
10. Nikith, B. V., Keerthan, N. K. S., Praneeth, M. S., Amrita, T.: Leaf disease detection and classification. Procedia Computer Science, vol. 218, pp. 291–300. Elsevier (2023).
11. Cagadas, D. O., Labajan, R. A.: Leaf-Based Cacao Diseases Classification Using Image Processing. Sci. Int. (Lahore) (2024).
12. Saleem, M. H., Potgieter, J., Arif, K. M.: Plant disease detection and classification by deep learning. Plants, vol. 8, no. 11, pp. 468. MDPI (2019).
13. Li, L., Zhang, S., Wang, B.: Plant disease detection and classification by deep learning—a review. IEEE Access, vol. 9, pp. 56683–56698. IEEE (2021).
14. Oo, Y. M., Htun, N. C.: Plant leaf disease detection and classification using image processing. International Journal of Research and Engineering, vol. 5, no. 9, pp. 516–523. Marwah Infotech (2018).
15. Dhingra, G., Kumar, V., Joshi, H. D.: Study of digital image processing techniques for leaf disease detection and classification. Multimedia Tools and Applications, vol. 77, pp. 19951–20000. Springer (2018).