International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Oct-24)

http://dx.doi.org/10.12785/ijcds/1601125

Leaf Condition Analysis Using Convolutional Neural Network and Vision Transformer

Wai-Chun Yong[1], Kok-Why Ng[1], Su-Cheng Haw[1], Palanichamy Naveen[1] and Seng-Beng Ng[2]

[1]*Faculty of Computing and Informatics, Multimedia University,63100 Cyberjaya, Selangor, Malaysia*
[2]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

Received 28 Feb. 2024, Revised 17 Jul. 2024, Accepted 2 Aug. 2024, Published 1 Oct. 2024

Abstract: Plants play an essential role to human survival, from being the primary source of oxygen emissions to being a vital supply of dietary ingredients. It keeps the ecosystem's general equilibrium, particularly in the food chain. Diseases will cause plants to deteriorate in quality. Many botanists and domain experts research various ways to prevent plants from getting infected and preserve their quality using computer vision and image processing integration on leaf images. The quality of the image collection provides a substantial value for the classification model in identifying leaf diseases. Nevertheless, the amount of leaf disease image dataset is very scarce. Since the performance of the models is determined on the overall quality of the dataset, this could compromise the predictive models. Besides, existing leaf disease detection programs do not provide an optimized user's experience. As a result, although customers may receive an excellent interactive features programme, the backend algorithm is not optimized. This problem may discourage users from applying the program to solve plant disease problems. In this paper, contrast boosting, sharpening, and image segmentation are used to create an unprocessed leaf disease image dataset. Through the use of a hybrid deep learning model that combines vision transformer and convolutional neural networks for classification, the algorithm can be optimized. The model performance is evaluated and compared with the other methods to ensure quality and usage compatibility in the plantation domain. The model training and validation performance is represented on graphs for better visualization .

Keywords: Plant Disease Identification, Deep Learning, Vision Transformer, Convolutional Neural Network

1. INTRODUCTION

Ensuring the equilibrium of the environment depends in large part on the presence of flora and wildlife. The way of life for all living beings on Earth is in danger if unanticipated events disrupt the entire ecosystem. However, when it comes to preserving the ecosystem, flora does not get as much credit and attention as wildlife.

Botanists have been adopting various approaches of implementation in monitoring plant condition. One of the approaches is to analyze the leaf of the plant to determine the diseases. Leaf exposes in the air where the rain and wind carry spores of the pathogen to the plant tissue and spread the disease. Diseases can deteriorate the general quality of the plant such as abnormal increase in tissue size, curling or twisting the leaves, defoliation, dwarfing and wilting to harm the plant. The common diseases are black spots, downy mildew, blight, powdery mildew, and canker.

Image processing and computer vision are fields of using computers to process digital images via algorithms and allow computers to draw information from images [1]. Some of the image processing techniques consist of image segmentation, classification and transformation. Nowadays, image processing is extensively used in many different areas, specifically when image data is highly accessible and in a vast amount of volume [2]. These image data can provide useful insights. However, there are some images of low resolution or quality that need to be processed, image processing is capable of solving these issues by accepting a mass image data set and learning from the data to draw patterns and information to process later on. With that, machine learning's algorithms are incorporated in image processing to enhance the overall performance.

Image classification algorithm is good for identifying the type of plant disease [3], [4]. The purpose of the image data set is to feed to the classification model for training. With optimized configuration, the model will be able to categorize the plant disease accurately with less time-consuming. However, there are a huge variety of algorithms that can be used for image classification. Choosing the right algorithm will help improving the accuracy and also optimizing the overall performance of the model.

The contributions of this paper are as follows:

E-mail address: waichun.yong01@gmail.com, kwng@mmu.edu.my, sucheng@mmu.edu.my

• Detailed reviews of some image processing models especially in Deep Learning field and its hybrid models.

• An optimized image classification algorithm for identifying the type of plant disease is proposed. With optimized configuration, the model can categorize the plant disease accurately with less time-consuming.

• Experimental evaluation on the proposed hybrid model with some existing models.

• A program interface is developed to display the leaf condition to identify the type of diseases and provide suitable cure or cultivation by analyzing the leaf condition.

Next section will review some existing literature. Section 3 will propose our algorithm. Experimental result will be discussed in Section 4. Section 5 will conclude this paper.

2. Literature Review

This section will feature the exploration of the latest research papers which are related to plant disease image processing techniques. Several prominent image processing methods will be discussed and analyzed in detail. Image segmentation is part of the image processing stages where certain parts of an image are cropped and fed to the classifier. Segmentation [5], [6] can ensure the classifier produces high accuracy and it will ease the classification process. Kurmi et al. [7] proposed a Random Sample Consensus (RANSAC) segmentation method on a PlantVillage dataset. This method comprised 3 stages. The first stage is to separate the leaf part from the background. Then, it limits the initial seed identification with the region growing algorithm. Next, boundary extraction and curve fitting with the RANSAC algorithm to crop the leaf part clearly. The method helped the classifier to provide an accuracy of 93.2%. Chouhan et al. [8] did a segmentation process using a Hybrid Neural Network that worked along with Superpixel clustering. Simple Linear Iterative Clustering (SLIC) was used for the superpixel clustering, and Adaptive Linear Neuron (ADALINE) was used for real-time de-noising. SLIC utilizes a metric calculation system by gaining the computed distance from every pixel to the center of the cluster. Moreover, Sathiya et al. [9] utilized a Multi-swarm Coyote Optimization (MSCO) algorithm for image segmentation. This method was based on the image's color information, borders, or parts to perform the separation process.

Feature extraction is part of the stages in processing a dataset to extract useful information so that the classification stage can perform better, and the result will be more accurate [10]. Conventional machine learning models do not have the ability to extract features by themselves. They heavily rely on feature extraction methods to boost the process and accuracy. Common image feature extraction algorithms are Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and features from accelerated segment test (FAST). Umamageswari et al. [11] proposed a feature extraction method using a fast Gray Level Co-occurrence Matrix (GLCM) model. GLCM was used to extract image texture descriptions and fast GLCM could solve the stringer processing resource problem. Khan et al. [12] also proposed the use of GLCM related model, which is an advanced form of HOG, Pyramid of HOG (PHOG) with GLCM to extract the diseased infected part. The PHOG provided a recurrence zone representation of the infected spots whereas the GLCM was used to compute the texture image feature.

Since its establishment, machine learning techniques have widely used in image classification. Machine learning models are amalgamations of mathematical formulations to form an algorithm that can recognise patterns. Some of the common models are linear regression, random forest, support vector machines, decision tree and XGBoost. Jackulin et al. [13] proposed a comparative study between some machine learning methods. The machine learning ideas proposed were Support Vector Machine and Random Forest. The authors also did a comparative analysis between the machine learning techniques with CNN on a plant disease dataset. Kirola et al. [14] proposed another comparative study with SVM, K-Nearest Neighbour (KNN), Random Forest, Logistic Regression and Naive Bayes on a plant disease dataset. They concluded that Random Forest achieved the best recognition accuracy. On the other hand, Generative Adversarial Network is a popular deep learning algorithm that can oversample data as the data it generated is very realistic to a point where humans have difficulty in recognising the authenticity of the data [1]. Nafi et al. [15] had proposed a GAN-based approach with a ResNet classifier to overcome the hurdle of class imbalance. Gandhi et al. [16] proposed another similar model which was using CNN and GANs as an augmentation purpose. GANs had been unstable, and it would produce outputs that hardly made sense. They used deep convolutional generative adversarial networks (DCGAN) which was a modified version, built for stable training. Abbas et al. [17] proposed another variant of GAN called the Conditional Generative Adversarial Network (C-GAN) with a pre-trained DenseNet121 model. C-GAN was used to produce synthetic tomato images and the images were fed to the classifier. C-GAN is a modified GAN model that comprised two adversarial network generators and discriminator model with four convolutional layers. Convolutional Neural Network (CNN) uses convolutional filtering layers to extract useful features. Some of the common CNN models are VGG, ResNet, DenseNet, MobileNet, Inception and AlexNet. CNN architecture can be varied depending on the purpose of optimizing the algorithms. Paymode et al. [18] implemented the Visual Geometry Group (VGG) transfer learning model. They applied the VGG16 model to the PlantVillage dataset with an amalgamation of different created real field images. The VGG16 model is an improved model with kernel-capacity filters along with eleven and five convolutional blocks that has a 3X3 kernel-capacity filter. Ali et al. [19] proposed the usage of conventional

CNN to identify the protein deficiency of grape leaves. The model consists of two 5X5 convolutional blocks and two 2X2 subsampling blocks as a feature selection layer with a classification layer of a fully connected network.

Aside from the CNN models, there are other related works proposed unconventional CNN models. Zhao et al. [20] proposed a CNN model that was based on Inception with residual structure that incorporates in an embedded modified convolutional block attention module (CBAM), they called it the RIC-Net network. This model performed well in the plant leaf disease classification. The authors applied the model on corn, potatoes, and tomatoes dataset. Traditional RI-Net networks are easily confused during the recognition of different kinds of leaf disease samples. The proposed CBAM solved the problem by detecting the location of the leaf disease from the image and enhancing the model detection quality. Wang et al. [21] suggested a Trilinear CNN (T-CNN) model with bilinear pooling. The T-CNN model has three CNN streams, one stream was used for area recognition tasks in picture detection and the rest of the two streams functioned to extract features for crop and sickness detection. The model was designed to focus more on the related features; hence the impact of irrelevant features was decreased, and the crop classification is more relevant to the real-world environment. Wang et al. [22] constructed a lightweight network architecture called MS-DNet. The model size is relatively small, but the computational speed is fast. The model is a DenseNet that consists of 3 layers of dense blocks that incorporate the SE module. It was incorporated to utilize the channel-wise attention of the model to study the inter-dependency features to maximize the reusing of the channels. Kaur et al. [23] proposed a modified Mask Region CNN (Mask R-CNN) that could perform segmentation and recognition as autonomous on a tomato plant leaf data set. The author expanded a lightweight "Region CNN (R-CNN)" to save memory usage and computational cost.

Likewise, Vision Transformer (ViT) is a newly introduced Deep Learning algorithm for image processing purposes in 2021. It starts by splitting the image into patches to be flattened. Next, lower dimensional linear embeddings are generated from the flattened patches and positional embeddings are appended. Furthermore, the transformer encoder will accept a strand of sequence as an input. The model is trained with image labels. Lastly, fine-tuning is done to achieve the optimized result. However, unlike CNN, ViT is generally harder to implement because of the complexity of its code and the architecture is composed with different types of layers. Thai et al. [24] utilized the model of ViT to classify cassava leaf diseases dataset. The model initially split the cassava leaf dataset into predetermined size proportions, the shape could either be 16X16 or 32X32. Next the patches were flattened and passed into the embedding vector with a unique token added in the beginning. All the embedded patches were processed in the Transformer Encoder and the output were classified by the Feed Forward Neural Network. Wang et al. [25] proposed a variant of the ViT model called Swin Transformer (SwinT) model. The model is a backbone-based network that depends on an enhanced SwinT and it is executed on a cucumber leaf diseases dataset. The proportion section of SwinT was upgraded using stepwise small proportions-embeddings for improving the performance of extracting the feature without adding the number of parameters. Jajja et al. [26] proposed a model called Compact Convolutional Transformer (CCT) on the AgriPK Dataset, a cotton leaf disease dataset. The model was created with a kernel size of 3X3 convolution layers, dual tokenizer and a stride size of 1. A ReLU activation function was used. Hirani et al. [27] proposed a comparative experiment with ViT, CNN and Transfer Learning. For the CNN model, the authors implemented a customized CNN model with three convolution blocks along with a ReLU activation function and three MaxPooling2D blocks.

Generally. hybrid models are more advanced and complicated, but their performance is more robust. Li et al. [28] proposed a hybrid model between ViT and CNN. They called it the ConvVit. They tested the model on a self-built kiwifruit disease dataset. The authors merged the convolutional structure and the transformer structure. The convolutional structure had a 3X3 depth-wise convolutional block connected to a layer norm block and a 1X1 point-wise convolutional block with a residual connection. Lu et al. [29], [30]constructed a hybrid model of Ghost convolutional and Transformer networks (GET) for identifying grape leaf dataset from a field. The authors made use of Ghost convolutional network as a backbone to produce featurettes with simple linear equations. Belay et al. [30] developed a hybrid model of CNN and Long Short-Term Memory (LSTM) on a chickpea images dataset. The CNN and LSTM were responsible to extract global level features and heavy spatial and timely features from the pictures to enhance the identification capability of the CNN's framework. Nandhini et al. [31] proposed a new episodal pictures classification hybrid model using Recurrent Neural Network (RNN) and CNN on a plant disease dataset. They called it the Gated-Recurrent CNN (G-RecConNN). The convolutional layers were used to extract high-level features through learning the spatial correlation between the image pixels. Bedi et al. [32] created a hybrid model with Convolutional Autoencoder (CAE) network and CNN on a peach plant leaf images dataset. The CAE network in this model was trained for dimensionality reduction while keeping the crucial features whereas the CNN was used for classification of the leaf condition.

3. Proposed Algorithm

Figure 1 shows the overall design of the theoretical framework. The algorithm starts with data acquisition in image format, then followed by image segmentation through cropping the leaf areas. The segmented leaf images undergo the image pre-processing stage which comprises of image sharpening, image contrast stretching and image resizing. The processed dataset is resampled using an under-sampling

method to balance the dataset. In addition, image augmentation is applied to increase data variation and balance the dataset. Next, the dataset is split and fed to the classification model to train, validate and test. After successfully training the model, a few leaf images are used to predict the plant and disease type. The program then displays the plant type and disease type. Lastly, it displays a summary of causes and possible treatments and ways of prevention. The following sections will explain each event of the algorithm.

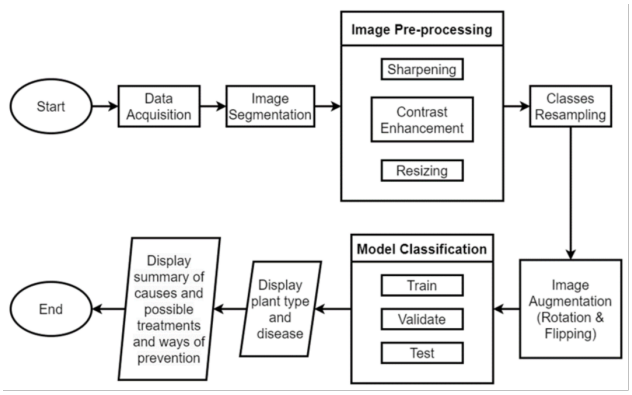


Figure 1. Theoretical framework of proposed algorithm

A. Data Acquisition

Data acquisition is very crucial because choosing the ideal dataset could affect the performance of the model. However, the biggest challenge of acquiring leaf diseases dataset is the quantity scarcity. The chosen dataset for this paper is PlantVillage dataset. It is a multiclass dataset, and the variation of the dataset is very wide. However, the dataset is very noisy, and it needs to be cleaned or denoised. PlantVillage dataset is good for image exploration, processing, analysis, and model evaluation because it is raw, original and diverse.

B. Image Segmentation

The proposed method of segmentation is color-based segmentation. Since the leaf regions are varied from green with slight yellow and brown as the disease spots. It is easy to perform image segmentation using color-based detection by detecting the color of the leaf region. The perk of using this technique is it is easy to implement with low computational time and complexity. Besides, the result obtained is also splendid. However, this method may be sensitive to detecting redundant objects with similar color range. Hence, it may wrongly segment the overall region.

C. Image Pre-Processing

This stage enhances the image quality to bring out the details and information for the model to perform better in predicting the leaves and diseases. From Figure 2, the first process is to sharpen the details of the leaf image. This method is useful for making the details and information appear more prominent. Hence, when performing the classification stage, the model may recognise the texture and feature easier. After sharpening, contrast stretching is performed to enhance the contrast of the leaf image. This technique is crucial for enhancing the clarity of the image by preserving the overall brightness. This method can bring out the leaf characteristic more by increasing the brightness difference between the leaf green region and the disease regions. In the end, image standardization is done by resizing all the images into the same size so that the model can accept all the images during the training phase. The size of the images is standardized to 256 X 256 respectively.

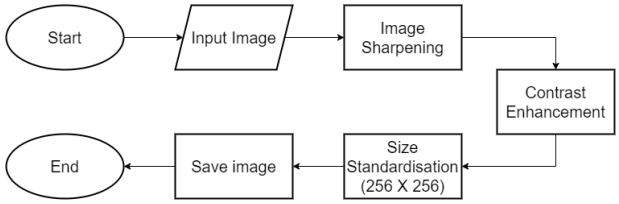


Figure 2. Image pre-processing stages

D. Classes Resampling

Figure 3 displays the total number of images in each class of the segmented and pre-processed image dataset. The dataset is imbalanced, and it biases towards the tomato yellow leaf curl virus class as it has the most number of images which is 5356. The specific class is resampled to approximately 2500 to 3000 images using under-sampling technique. After that, the dataset is split into train, validate, and test folders.

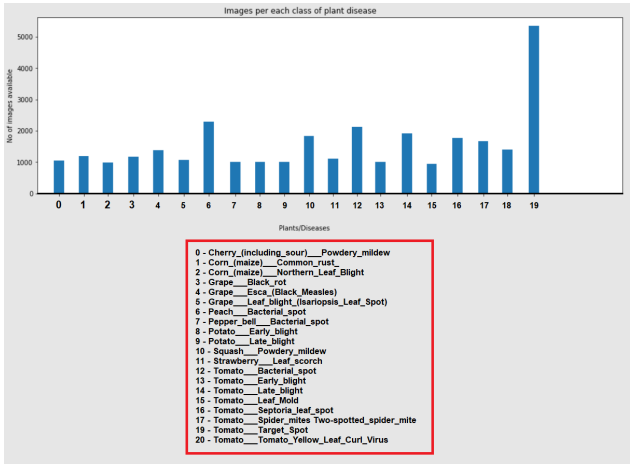


Figure 3. Number of every selected leaf picture in a class

E. Image Augmentation (Rotation and Flipping)

For classes with less images, image augmentation is done to expand the size of the classes by adding more variety into the classes. The methods of performing rotation 90°, 180°, 270° and mirror flipping are done. The augmented images are randomly chosen and added to

their classes respectively to achieve a total amount of approximately 2000 to 2500 images. These processes are useful for balancing the overall dataset and it can increase the performance of the classification model to identify the images more accurately in different angles and sides.

F. Model Classification)

Many computer vision applications have been adopting the use of CNN in recognising daily images. On the other hand, ViT was introduced back in 2021. The model was inspired by the Transformer which is used for natural language processing (NLP) tasks, instead the ViT breaks the image into patches and linear embedding is applied on the patches before feeding it to the transformer encoder. Over the years, ViT has been tested and compared with CNN and many results have shown that ViT has surpassed multiple CNN models at accuracy and performance. As such, we adopt the approach of combining these two models. In addition, both CNN and ViT perform well at extracting local and global features from images: CNN is assigned to extract the local features of the leaf images like the histogram-based features of the textures which will be good for recognising diseases whereas, ViT will be assigned to extract global features of the whole image in order to identify and detect the existence of the object so that the leaf types can be differentiated accurately. The model then goes through the training, validation and testing phase in order to evaluate the overall accuracy and performance.

G. Summary of Causes and Possible Treatments and Ways of Prevention

The application displays a summary of the causes, the possible treatments to attempt and the possible ways to prevent the disease from infecting and spreading off. The causes mainly discuss what are the bacteria infection and their favorable environment condition. Moreover, the possible treatments and ways of prevention are summarizing the possible factors that can ease the process of spreading, and bacteria from growing rapidly. It also discusses possible ingredients of fungicides and pesticides that could be effective for killing the pests and bacteria.

4. Results and Discussion

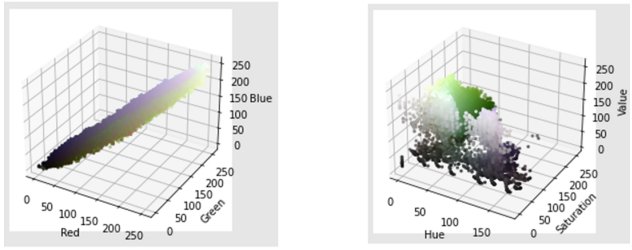
The PlantVillage dataset consists of 38 classes and 54272 images in total (see Figure 4) which can be categorized into 14 classes of unique plants. These categories are apple, blueberry, cherry (including sour), corn (maize), grape, orange, peach, pepper bell, potato, raspberry, soybean, squash, strawberry and tomato. In terms of disease, there is 26 unique disease classes: apple scab, black rot, cedar apple rust, powdery mildew, cercospora leaf spot gray leaf spot, common rust, northern leaf blight, black rot, esca (black measles), leaf blight (isariopsis leaf spot), haunglongbing (citrus greening), bacterial spot, bacterial spot, early blight, late blight, powdery mildew, leaf scorch, bacterial spot, early blight, late blight, leaf mold, septoria leaf spot, spider mites two-spotted spider mite, target spot, tomato mosaic virus and tomato yellow leaf curl virus.

However, not all the 38 classes are fully used due to the dataset is severely imbalanced in general. Although data balancing processes such as data resampling and data augmentation are carried out, the difference between the maximum leaf images and the minimum leaf images is significant. Data under-sampling for the classes that have a high number of images may cause the accuracy drops and hinders the model overall robustness. In the end, it can affect the model evaluation process. Besides, processing all the dataset will consume a lot of time especially in the model training phase. This paper requires analysing the texture, colour, disease region and the shape of the leaves. Too much shrinkage may lose the information of the images. Eventually, it disrupts the model's performance on predicting the leaves and diseases accurately. As such, 20 classes of the leaf images has been chosen.

Figure 4 shows the sample input image for segmentation. The segmentation process converts the input image color space from BGR to RGB so that the color pixel is aligned with the color in the original image file. Next, the image color space is converted again from RGB to HSV as HSV describes color more relatable to how human eyes perceive colors. Besides, HSV also defines color by separating color information from luminance.

Figure 4. Sample of origianl leaf images

Figure 5(a) shows the 3D scatter plot of the leaf image in RGB whereas Figure 5(b) is in HSV. As observable, the RGB scatterplot is clumped together across the red, green and blue values making the color difficult to distinguish. On the contrary, the HSV scatterplot is easier to recognise each color as they are clustered in different groups without mixing.



(a) (b)

Figure 5. (a) Leaf image is RGB scatter plot; (b) Right image is HSV scatter plot

After converting the color spaces, the leaf is segmented

by detecting the colors of green, brown, and yellow. Green represents the natural green pigment part of the leaf produced by the chlorophylls, brown and yellow represent the potential disease spots on the leaf.

Figure 6(a) shows the result of the mask after segmenting the leaf region based on the respective color ranges. The outlook of the mask looks almost identical to the original image leaf shape. However, some post-processing must be done because of the holes inside of the mask. This issue is solved by filling the contours which are the boundary of the joined objects. Figure 6(b) shows the result. After filling the holes, a morphological operation of closing is implemented to smoothen the overall edges and reduce the noise. The result is shown in Figure 6(c), which looks smoother and less noisy than Figure 6(b).

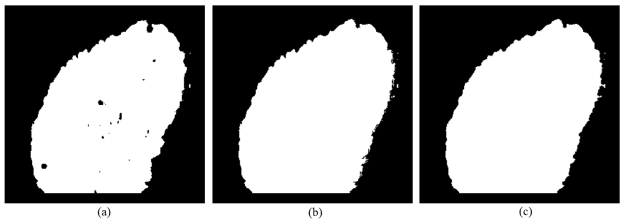


Figure 6. (a) Segmented leaf mask (b) Leaf mask after holes filling (c) Leaf mask after closing morphological operation.

Figure 7 shows the result of before and after performing image segmentation. The leaf outline of Figure 7(b) looks identical to Figure 7(a). This method is fast in computational speed, and it does not require high processing power. Figure 8(b) shows the sharpened image looks more clearer and sharper as compared to Figure 8(a). The leaf venation and disease spots are noticeably prominent in Figure 8(b).

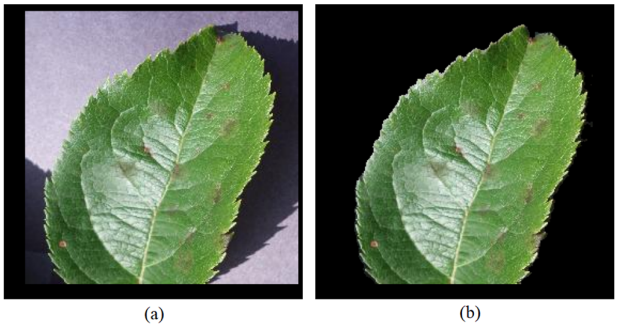


Figure 7. (a) Original leaf image (b) Segmented leaf image

Data augmentation is done to expand the size of the necessary classes. Figure 9 shows the way of rotating the image with transformation. After that, random numbers of augmented images are selected and moved to their respective classes to meet the requirement of having approximately 2000 to 3000 of images in each class. The first image is the original image, followed by the other rotated and flipped images. The computation equations are then applied.

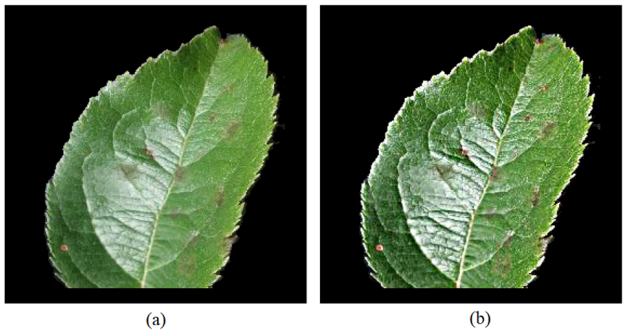


Figure 8. (a) Leaf image before sharpening (b) Leaf image after sharpening.

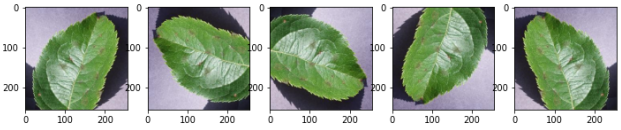


Figure 9. Result of Image Augmentation

Equation (1) shows the convolution filtering formula that is used to compute the feature map values. F is equivalent to the input image and H denotes the kernel whereas i and j represent the indices of row and column. The filtering is done by multiplying the input image, window by window with the kernel in pairs. Equation (2) is the min-max normalization and Equation (3) is the coefficient of variance.

$$G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} H[u,v]F[i-u, j-v] \quad \textbf{(1)}$$

$$x_{normalized} = \frac{x - m}{x_{max} - x_{min}} \quad \textbf{(2)}$$

$$x_{normalized} = \frac{x - m}{s} \quad \textbf{(3)}$$

The hybrid model implementation is inspired by and adopted from the proposed method of Li et al. (2022). They used the same method of CNN-ViT hybrid as well. However, the domain implementation is different from this paper. They mainly performed classification on the strip steel surface fault. Figure 10 shows the proposed model architecture.

The accuracy calculation is computed using the Equation (4). Equation (5) is the formal and expanded version. Figure 11(a) shows the graph of accuracy trend over epochs. The model fitted perfectly as the training accuracy and the validation accuracy have the same patterns and trends. This denotes that there is not a sign of overfitting and underfitting which shows the robustness of the model and the dataset. Figure 11(b) shows the graph of the loss

trend over epochs. It shows that the training loss overall pattern is the same as the validation loss pattern.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FP} \quad (4)$$

$$Accuracy = \frac{Numberof CoherentPrediction}{TotalNumberof Prediction} \quad (5)$$

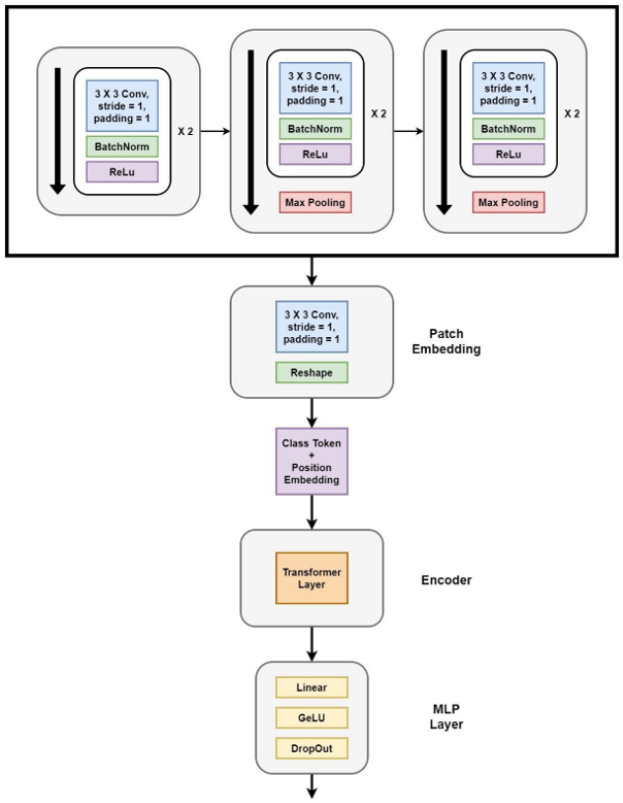
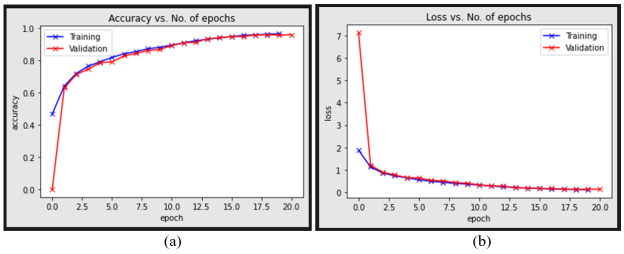


Figure 10. Proposed CNN-ViT hybrid model architecture



(a) (b)

Figure 11. (a) Accuracy over epochs graph (b) Loss over epochs graph

Four models are chosen to perform the efficiency comparison in terms of their accuracy, speed, and memory usage. The models are the proposed CNN+ViT hybrid model, CNN baseline model, ViT baseline model, and ResNet model. CNN and ViT baseline models are used to compare the difference between before and after hybrid. On the other hand, ResNet is used for comparison because ResNet has been proven to be the superior CNN model. It outperforms some of the latest deep learning models until today. ResNet is still very much relevant and not obsolete when it comes to accuracy, speed and power. In this case, ResNet-50 is used because we aimed to have a lighter-weight model instead of a heavier-weight. ResNet-50 can perform faster and consumes a lesser amount of memory usage although it may sacrifice a little accuracy, but it is still a very robust model when it comes to accuracy. All the models are trained under the same engine and environment which is the Google Colab's T4 GPU and the environment has 12.7 GB system memory, 15.0 GB GPU memory and 78.2 GB Disk space. The models are also hyperparameter tuned with a grid-searching method. The concept of hyperparameters is about the values used by the deep learning models to achieve certain performance of the models. A model with fine-tuned hyperparameters can obtain the most optimized performance. Examples of hyperparameters are number of epochs, dataset sizes, number of layers and the learning rate. However, some of the model's parameters cannot be tuned until the most optimized version because of the limitation of the system's engines. The system will crash due to certain number of epochs or image sizes. Below is the comparison of the models trained with 50x50 and 550x550 images dataset.

Table I and Table II show the comparative analysis among all the models that trained with 50x50 and 550x550 images dataset. The baseline CNN model performs the best when it comes to accuracy, time taken and the memory usage. The proposed hybrid CNN+ViT model comes second. Although it shows that CNN+ViT hybrid is not outperformed from CNN, the proposed model still excels as compared to other state-of-the-art models. The ViT architecture can perform efficiently when the dataset size is large. For the ResNet-50 model, it is the lightest weight version and can train and test on the 50x50 images dataset, but the memory usage surges very high at a point where it almost runs out of memory; whereas for the 550x550 images dataset, it is unable to show any testing result because of huge memory it requires. ResNet-50 may be able to outperform the proposed models (baseline and hybrid) but the amount of memory and time taken it requires are not worth the effort for such a light-weight dataset. Therefore, the proposed models produce good results and are able to fulfil the light-weight model's idea. The reason why the proposed hybrid model has the most balance performance out of all of the models is mainly because its architecture. In terms of speed and computation complexity, the hybrid model consists of two of the most basic versions from their own model types. The CNN that is used in basically an alternative version of VGG with some customizations, VGG model is known for its light-weighted nature. On the other hand, we also have the most basic version of ViT model which consist of the basic architecture of the model without any additional layers like the Swin Transformer

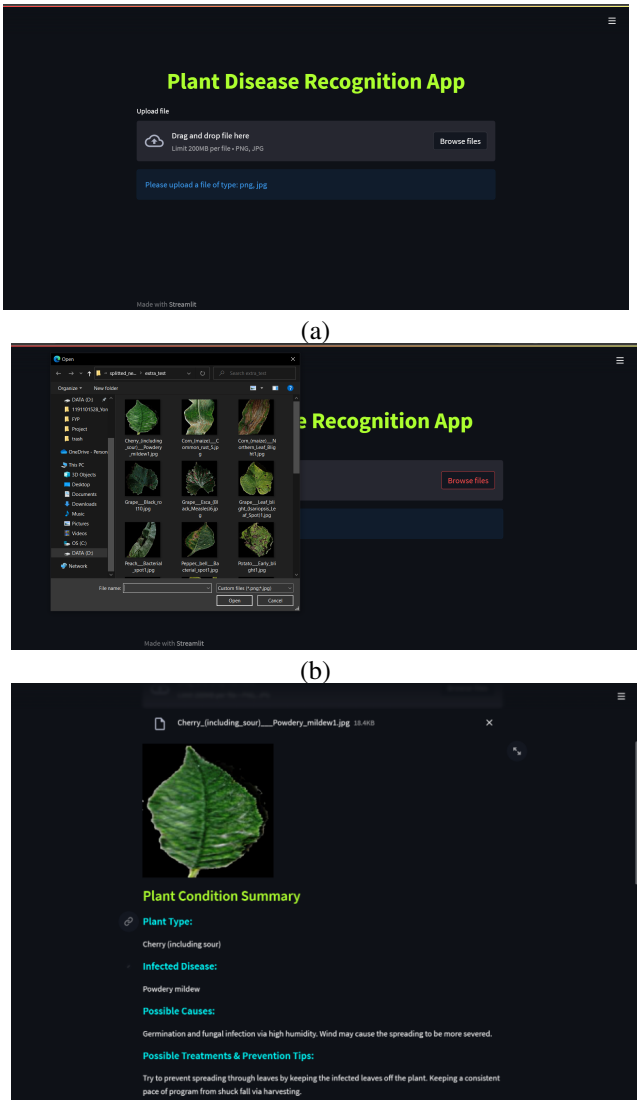


Figure 12. (a) Main webpage (b) File upload process (c) Summary display

or BERT. Therefore the speed and memory usage of the proposed hybrid model are on par with the basic CNN and ViT model because of both model's light-weighted nature. On the other hand, ResNet contains the residual layers also known as the Res layer and these layers actually add more computational steps to the model and it results in heavier cost of algorithm. Furthermore, combining two models will also enhance the accuracy of the model because CNN will first train the datasets and generate an output of feature maps to the ViT and it will train once more, increasing the predicting accuracy percentage. This is also why ResNet has high predicting accuracy because ResNet has many Residual layers that can analyze and compute the feature maps of the image dataset. The residual layers consist of many other components that are used in extracting the feature maps of the image. The components are mentioned in previous chapters.

Although the proposed hybrid model was the most balanced out of all the compared models. The process of tuning and optimizing the model's performance can also be quite a hassle for most of the time since the hybrid model is about combining two models, so the hyperparameters will also increase. Both of the models have to be tuned to the best with methods like grid searching and it will take a very long time to train a model on large datasets since the datasets are all images, the tuning time will take over a day. Besides, a hybrid model is also harder to be maintained compared to standalone models since there are two models to be maintained.

Figure 12 shows some sample screenshots of the web application. Figure 12(a) illustrates the main page of the program. Users can click the upload files button and they are given the options to upload JPEG/JPG and PNG file format. Figure 12(b) displays the process of uploading the file and users can choose to perform drag and drop options as well. In the end, Figure 12(c) shows that the program displays a summary by identifying the leaf and disease types. It also displays the possible causes and possible treatments and prevention tips. The web application is very user-friendly and it also get straight to the point with main information that should be conveyed as opposed to other diagnostic application that generate many unwanted information which gives the users a hard time to perform analysis and examination.

5. Conclusion

This paper presented the hybrid classification models using CNN and ViT. Both can extract feature maps in a detailed approach. CNN is used to extract bottom level features whereas ViT is used to extract top level features and classify the images. The model achieved an accuracy running from 95% to 96% which shows a good result in terms of speed, memory usage and also accuracy. In addition, the model is deployed on a web interface that can identify leaf and disease types along with a summary of the leaf information.

In the future, different hybrid models should be tried out in order to find out the most optimal models. Besides, a baseline model should be evaluated first before performing hybrid combination in order to assess the need of performing hybrid.

References

[1] M. Y. Xin, L. W. Ang, and S. Palaniappan, "A data augmented method for plant disease leaf image recognition based on enhanced gan model network," *Journal of Informatics and Web Engineering*, vol. 2, no. 1, 2023.

[2] P.-W. Chin, K.-W. Ng, and N. Palanichamy, "Plant disease detection and classification using deep learning methods: A comparison study," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 155–168, Feb 2024.

TABLE I. Comparison of models trained with 50x50 images dataset

Models	Accuracy (%)	Speed (Seconds)	Memory Usage
CNN + ViT Hybrid	93.80%	7.0184	System ram: 5.0/12.7GB GPU ram: 3.3/15.0GB Disk: 26.1/78.2GB
CNN	93.44%	4.8878	System ram: 4.2/12.7GB GPU ram: 1.2/15.0GB Disk: 26.1/78.2GB
ViT	75.92%	5.6662	System ram: 4.8/12.7GB GPU ram: 3.7/15.0GB Disk: 26.1/78.2GB
ResNet-50	89.23%	8.7826	System ram: 12.1/12.7GB GPU ram: 6.0/15.0GB Disk: 26.1/78.2GB

TABLE II. Comparison of models trained with 550x550 images dataset

Models	Accuracy (%)	Speed (Seconds)	Memory Usage
CNN + ViT Hybrid	97.19%	24.2176	System ram: 5.1/12.7GB GPU ram: 7.8/15.0GB Disk: 26.1/78.2GB
CNN	97.24%	23.0851	System ram: 4.7/12.7GB GPU ram: 4.8/15.0GB Disk: 26.1/78.2GB
ViT	90.45%	24.0231	System ram: 4.6/12.7GB GPU ram: 9.6/15.0GB Disk: 26.1/78.2GB
ResNet-50	Not Available (out of memory) %	Not Available (out of memory)	System ram: 12.1/12.7GB GPU ram: 6.0/15.0GB Disk: 26.1/78.2GB

[3] S. S. Gaikwad, S. S. Rumma, and M. Hangarge, "Fungi classification using convolution neural network," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 10, 2021.

[4] K. Kavitha and S. Naveena, "Deep learning framework for identification of leaf diseases in native plants of tamil nadu geographical region," in *2023 International Conference on Computer Communication and Informatics, ICCCI 2023*, 2023.

[5] Y.-S. Leow, K.-W. Ng, Y.-J. Yoong, and S.-B. Ng, "Sickle cell segmentation and classification for thalassemia aid diagnosis," *F1000Res*, vol. 10, p. 1185, Nov 2021.

[6] K. W. Ng, X. H. Lam, Y. J. Yoong, and S. B. Ng, "Wbc-based segmentation and classification on microscopic images: A minor improvement," *F1000Res*, vol. 10, 2021.

[7] Y. Kurmi, S. Gangwar, D. Agrawal, S. Kumar, and H. S. Srivastava, "Leaf image analysis-based crop diseases classification," *Signal Image Video Process*, vol. 15, no. 3, 2021.

[8] S. S. Chouhan, U. P. Singh, and S. Jain, "Web facilitated anthracnose disease segmentation from the leaf of mango tree using radial basis function (rbf) neural network," *Wirel Pers Commun*, vol. 113, no. 2, 2020.

[9] V. Sathiya, M. S. Josephine, and V. Jeyabalaraja, "An automatic classification and early disease detection technique for herbs plant," *Computers and Electrical Engineering*, vol. 100, 2022.

[10] M. Xin, L. W. Ang, and S. Palaniappan, "A multi-scale feature attention image recognition algorithm," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, 2023.

[11] A. Umamageswari, N. Bharathiraja, and D. S. Irene, "A novel fuzzy c-means based chameleon swarm algorithm for segmentation and progressive neural architecture search for plant disease classification," *ICT Express*, vol. 9, no. 2, 2023.

[12] S. Khan and M. Narvekar, "Disorder detection in tomato plant using deep learning," 2020.

[13] C. Jackulin and S. Murugavalli, "A comprehensive review on detection of plant disease using machine learning and deep learning approaches," *Measurement: Sensors*, vol. 24, 2022.

[14] M. Kirola, K. Joshi, S. Chaudhary, N. Singh, H. Anandaram, and A. Gupta, "Plants diseases prediction framework: A image-based system using deep learning," in *Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022*, 2022.

[15] N. M. Nafi and W. H. Hsu, "Addressing class imbalance in image-based plant disease detection: Deep generative vs. sampling-based

approaches," in *International Conference on Systems, Signals, and Image Processing*, 2020.

[16] R. Gandhi, S. Nimbalkar, N. Yelamanchili, and S. Ponkshe, "Plant disease detection using cnns and gans as an augmentative approach," in *2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018*, 2018.

[17] A. Abbas, S. Jain, M. Gour, and S. Vankudothu, "Tomato plant disease detection using transfer learning with c-gan synthetic images," *Comput Electron Agric*, vol. 187, 2021.

[18] A. S. Paymode, S. P. Magar, and V. B. Malode, "Tomato leaf disease detection and classification using convolution neural network," in *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, 2021.

[19] A. Ali, S. Ali, M. Husnain, M. M. Saad Missen, A. Samad, and M. Khan, "Detection of deficiency of nutrients in grape leaves using deep network," *Math Probl Eng*, vol. 2022, 2022.

[20] Y. Zhao, C. Sun, X. Xu, and J. Chen, "Ric-net: A plant disease classification model based on the fusion of inception and residual structure and embedded attention mechanism," *Comput Electron Agric*, vol. 193, 2022.

[21] D. Wang, J. Wang, W. Li, and P. Guan, "T-cnn: Trilinear convolutional neural networks model for visual detection of plant diseases," *Comput Electron Agric*, vol. 190, 2021.

[22] D. Wang, J. Wang, Z. Ren, and W. Li, "Dhbp: A dual-stream hierarchical bilinear pooling model for plant disease multi-task classification," *Computers and Electronics in Agriculture*, vol. 195, 2022.

[23] P. Kaur *et al.*, "Recognition of leaf disease using hybrid convolutional neural network by applying feature reduction," *Sensors*, vol. 22, no. 2, 2022.

[24] H. T. Thai, N. Y. Tran-Van, and K. H. Le, "Artificial cognition for early leaf disease detection using vision transformers," in *International Conference on Advanced Technologies for Communications*, 2021.

[25] F. Wang *et al.*, "Practical cucumber leaf disease recognition using improved swin transformer and small sample size," *Comput Electron Agric*, vol. 199, 2022.

[26] A. I. Jajja *et al.*, "Compact convolutional transformer (cct)-based approach for whitefly attack detection in cotton crops," *Agriculture (Switzerland)*, vol. 12, no. 10, 2022.

[27] E. Hirani, V. Magotra, J. Jain, and P. Bide, "Plant disease detection using deep learning," in *2021 6th International Conference for Convergence in Technology (I2CT)*. IEEE, 2021, pp. 1–4.

[28] X. Li *et al.*, "Transformer helps identify kiwifruit diseases in complex natural environments," *Comput Electron Agric*, vol. 200, p. 107258, Sep 2022.

[29] X. Lu *et al.*, "A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1755–1767, May 2022.

[30] A. J. Belay *et al.*, "Development of a chickpea disease detection and classification model using deep learning," *Inform Med Unlocked*, vol. 31, p. 100970, 2022.

[31] M. Nandhini, K. U. Kala, M. Thangadarshini, and S. M. Verma, "Deep learning model of sequential image classifier for crop disease detection in plantain tree cultivation," *Comput Electron Agric*, vol. 197, p. 106915, Jun 2022.

[32] P. Bedi and P. Gole, "Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network," *Artificial Intelligence in Agriculture*, vol. 5, 2021.

Yong Wai Chun graduated in 2023 with a Bachelor of Science in Computer Science, specializing in Data Science, from Multimedia University. Throughout his university years, he developed a strong interest in Artificial Intelligence and Big Data, excelling in these areas and completing numerous significant projects. His passion for other IT fields, such as Operating Systems and Computer Networking, has paved the way for his current role as a professional 5G Core Network Engineer. He can be contacted at email: waichun.yong01@gmail.com

Kok-Why Ng is an Associate Professor at Faculty of Computing and Informatics, Multimedia University. His research interests include Image Processing, Computer Vision, Chatbot, Recommender System, Computer Graphics etc. He leads several funded projects on Microscopic Image Processing, Modelling and Rendering, and Recommender Systems. He can be contacted at email: kwng@mmu.edu.my.

Su-Cheng Haw is Professor at Faculty of Computing and Informatics, Multimedia University, where she leads several funded projects on the XML databases. Her research interests include XML databases, query optimization, data modeling, semantic web, and recommender system. She is also the chairperson for Center for Web Engineering (CWE) and chief editor for Journal of Informatics and Web Engineering (JIWE). She can be contacted at email: sucheng@mmu.edu.my.

Palanichamy Naveen joined the Faculty of Computing and Informatics, Multimedia University after receiving Ph.D from Curtin University, Malaysia. She received her Bachelor of Engineering (CSE) and Master of Engineering (CSE) from Anna University, India. Her research interests include Smart Grid, Cloud Computing, Machine Learning, Deep Learning and Recommender system. She is involved in multiple research projects funded by Multimedia University. She can be contacted at email: p.naveen@mmu.edu.my.

Seng-Beng Ng received his B.Sc., M.Sc. and PhD in Computer Science from Universiti Putra Malaysia (UPM), in 2004, 2007 and 2015 respectively. He is currently a senior lecture in the Faculty of Computer Science and Information Technology, UPM and a member of the Computer Graphics, Vision and Visualisation (CGV2) research group. His research interest is point cloud processing, image-based modelling, and geometrical reverse engineering. He can be contacted at email: ngsengbeng@upm.edu.my.