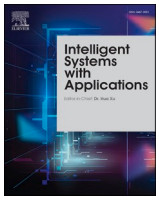
Intelligent Systems with Applications 26 (2025) 200490

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



A multi-class hybrid variational autoencoder and vision transformer model for enhanced plant disease identification

Folasade Olubusola Isinkaye [a,*], Michael Olusoji Olusanya [a,*], Ayobami Andronicus Akinyelu [b]

[a] Department of Computer Science and Information Technology, Sol Plaatje University, Kimberley, 8301, South Africa
[b] Department of Computer Science and Informatics, University of the Free State, P. Bag X13, Phuthaditjhaba, 9866, South Africa

ARTICLE INFO

Keywords:
Agriculture
Crop management
Deep learning
Food security
Machine learning
Plant diseases

ABSTRACT

Agriculture is considered as the propeller of economic growth as it accounts for 6.4 % of global gross domestic product (GDP) and in low-income countries, it can account for more than 25 % of GDP. Plants supply more than 80 % of the food consumed by humans and are the main source of nutrition for animals. Plant diseases pose a major risk to global food security as they account for losses of between 10 to 30 % of the global harvest every year. Deep learning techniques like convolutional neural networks successfully identify image-based diseases but struggle with capturing long-range contextual information. This makes them less robust in noisy or high-resolution images. Their high computational and memory demands also limit scalability for large datasets. To overcome these issues, we propose a hybrid model with the potential to combine Variational Autoencoders and Vision Transformers for enhanced accuracy and robustness of plant disease classification. Variational Autoencoder reduces image dimensionality while preserving essential features, and Vision Transformer captures global relationships to enhance accuracy and scalability, especially in multi-class disease classification. The experiment used images of corn, potato, and tomato plant leaves from the publicly available PlantVillage dataset. On-the-fly data augmentation was applied to further increase the robustness of the model. The proposed model achieved a classification accuracy of 93.2 %. This technique provides a reliable and efficient solution for identifying multiple plant diseases across various crops. It enhances agricultural productivity and supports food security efforts.

1. Introduction

There is a direct relationship between crop health and crop productivity. Consequently, enhanced plant health will result in increased crop productivity. Similarly, more than 80 % of the food that keeps humans alive comes from plants, and they are also the main source of nutrition for livestock. Therefore, it is very crucial to monitor plant health and identify disease at an early stage to limit its spread. Also, due to the crucial role that plants play in maintaining public health, the United Nations declared 2020 to be the International Year of Plant Health (IYPH) to raise public awareness of plant diseases and their underlying significant social impacts (Rizzo, Lichtveld, Mazet, Togami, & Miller, 2021).

Plant diseases cause severe crop damage, which reduces plant yields and increases food scarcity. This directly affects agricultural productivity and threatens food security globally (Muluneh, 2021). To enhance crop yield and optimize resource use in agriculture, there is a need for early and accurate disease identification for managing plant health effectively (Guo et al., 2020). However, traditional approaches solely depend on human knowledge, and are often labor-intensive, time-consuming, and prone to errors (Sajitha et al., 2024; Ahmed et al., 2024). These approaches are impractical and infeasible in large-scale farming environments as they could lead to inaccurate plant disease identification due to potential biases in decision-making (Orchi et al., 2021). Similarly, the increasing complexity of managing multiple crops and diseases simultaneously further compounds the difficulty of using these traditional methods. Hence, the necessity for automated, robust solutions that can efficiently identify plant diseases in complex agricultural settings (Ngugi et al., 2021).

Deep learning (DL) plant disease detection systems have the capacity to improve the management and control of plant diseases, which can contribute to more sustainable agricultural practices. Recently, DL techniques such as convolutional Neural networks (CNNs) have been successfully used to identify image-based diseases with reasonable performance. However, they face difficulties in multi-class disease classification due to the variability in disease symptoms, environmental

* Corresponding authors.
E-mail addresses: folasade.isinkaye@spu.ac.za (F.O. Isinkaye), michael.olusanya@spu.ac.za (M.O. Olusanya).

https://doi.org/10.1016/j.iswa.2025.200490
Received 17 October 2024; Received in revised form 8 January 2025; Accepted 7 February 2025
Available online 19 February 2025

conditions, and the need to differentiate between similar diseases across various plants. Their focus on local image features limits their ability to capture broader contextual information (Khan et al., 2023). Also, they are computationally demanding. This also makes it hard for them to consistently identify complex disease patterns and achieve high accuracy and robustness under such conditions (Mauricio, Domingues & Bernardino, 2023).

Recent advances in DL have introduced models such as Variational Autoencoder (VAE) (Bilodeau et al., 2022) and Vision Transformers (ViT) (Mauricio et al., 2023). VAEs are powerful generative models known for their potential to improve different components of machine learning tasks (Harshvardhan et al., 2020; Anstine & Isayev, 2023). They can be effectively used for tasks such as dimensionality reduction (Mahmud et al., 2020), data augmentation (Elbatta et al., 2021; Islam et al., 2021), feature extraction (Yao et al., 2019; Ma et al., 2020), and de-noising (Jung et al., 2018; Prakash, Krull & Jug, 2020; Xia et al., 2023) which can jointly contribute to enhancing the classification accuracy of the models they are integrated into (Arsenovic et al., 2019). Fre example, Tao et al. (2024) investigated quantized iterative learning control (ILC) with encoding-decoding mechanisms to reduce network problems and mitigate quantization errors in communication-constrained systems. Inspired by its parallels to Variational Autoencoders (VAEs) in handling resource limitations and uncertainties, this study implements the VAE-based technique by harnessing its potential to handle variability and constraints. This technique is very relevant in plant disease classification in agriculture, where challenges such as diverse imaging conditions and limited computational resources demand robust and efficient solutions. Similarly, Vision Transformers (ViT) are a type of neural network that uses self-attention mechanisms for image classification (Mao et al., 2022), object detection (Li et al., 2022) and semantic image segmentation tasks (Thisanke et al., 2023). Vision Transformer (ViT) architectures have been shown to outperform CNNs in classification tasks by effectively capturing global dependencies within an image, a key advantage over CNNs which tend to focus on local features (Khan et al., 2023).

Therefore, this study proposes a multi-class hybrid VAE and ViT model for enhanced plant disease identification by harnessing the potential of VAEs and ViTs to overcome the limitations of CNN. Specifically, VAEs will be used to extract high-quality features from plant leaf images, while ViTs will handle the classification task by capturing both local and global relationships within the image. This ability is particularly useful in multi-class classification, where different disease patterns may require attention to both local and global features. By capturing these relationships, ViTs enhance the model's accuracy in distinguishing between multiple diseases. So, the combined model is expected to enhance the accuracy and the robustness of plant disease detection model. This is also anticipated to contribute to better crop management and reduce the impact of plant diseases on agricultural productivity.

The rest of the paper is structures as follows: Section II provides a review of related work on plant disease classification utilizing various machine learning models. Section III details the proposed methodology for the hybrid model. Section IV presents and discusses the experimental results, and finally, Section V provides the conclusion and future research directions.

2. Related work

Manual inspection was the key technique for identifying plant diseases before ML and DL algorithms came into existence (Kotwal et al., 2023). These techniques are inefficient because they are susceptible to errors when dealing with multiple diseases. Likewise, there could be a reduction in the accuracy of the systems due to factors such as environmental conditions and subjective judgment. Automated systems based on ML and DL have the potential to enhance the accuracy and robustness of plant diseases identification systems (Kotwal et al., 2023).

To identify and track plant diseases, conventional approaches like support vector machines (SVMs), random forests (RF), Decision Trees (DT), and K-Nearest Neighbors (KNN) and other similar approaches have been used. For instance, in the work of Panigrahi et al. (2020), the study examined the performances of Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), SVM, and RF for maize plant disease detection. The result showed that RF algorithm has the highest accuracy of 79.23 % as compared to the rest of the classification algorithms. Three different types of grape leaf diseases were identified in the work of Kaur et al. (2019). Fractional-order Zernike Moments (FZM) was used for image feature extraction and SVM for classification with an accuracy of 97.3 %. In another study conducted by Yousuf and Khan (2021), ensemble ML models, such as RF and KNN were employed to detect plant disease with an accuracy of 96.0 % on three plant leaves. Limitations such as handling large amounts of high-dimensional data, problems of handling multiclass tasks, and feature extraction make these techniques less effective in some applications. CNNs have been acclaimed to be generally superior to conventional ML approaches when it comes to these tasks (Khan et al., 2023). Shelar et al. (2022) proposed a CNN plant leaf diseases system using 15 classes of plant diseases. They reported an accuracy of 88.8 %.

In plant disease detection, VAEs have been successfully utilized for feature extraction and dimensionality reduction. The VAE encoder captures the latent features of leaf images, which are then passed to classifiers for disease recognition. For instance, Majikumna et al. (2024) integrates focal loss variational autoencoders with Grey Wolf Optimization-based CNN for olive leaf disease classification. This approach optimizes the extracted latent features to enhance the accuracy of the classifier. It reduces irrelevant information and emphasizes critical patterns. Similarly, Nagachandrika et al. (2024) employs multi-scale feature fusion in a deep adaptive network. It combines the dimensionality reduction of VAE with advanced CNNs to identify plant diseases across varying scales and complexities.

The application of Vision Transformers (ViTs) in image classification has also significantly advanced state-of-the-art approaches in various domains, such as plant disease detection. Unlike traditional CNNs, ViTs use self-attention mechanisms to process a complete image holistically. This aids in capturing global contextual information and subtle patterns critical for distinguishing between visually similar disease symptoms. For example, De Silva and Brown (2023) showed the superior performance of ViT in multispectral plant disease detection by employing the spectral variations for accurate classification, Barman et al. (2024) highlighted the ability of ViTs to achieve high accuracy in smart agriculture applications, even in scenarios with complex or noisy datasets.

Furthermore, several studies have successfully applied VAEs or ViTs as individual components within hybrid models across various domains. For instance, Ahmed, Barua, Fahim-Ul-Islam and Chakrabarty (2024) introduced a novel approach for rice disease detection using CNN and transformer architecture. They modified a BEIT (Bidirectional Encoder representation from Image Transformers) architecture by reducing the number of parameters from 85 million to 65 million. The lightweight technique was evaluated on a rice plant dataset consisting of 5503 healthy and non-healthy leaf images, and it achieved a precision, recall, and F1 score of 92 %, 91 %, and 91 %, respectively. In another study, Faisal et al. (2023) proposed a hybrid feature fusion technique for detecting coffee leaf diseases. They introduced seven hybrid models to extract features from images using Swin Transformers, VAE, and MobileNetV3. MobileNetV3 was used to extract local features, while Swin Transformers was used to extract high-level features. The extracted features were combined and used to train MobileNetV3. The technique was evaluated on a coffee leaf dataset, and it achieved a classification accuracy of 84.29 %. In Omondi and Olwal (2023), VAEs was used to improve MIMO signal detection by capturing the underlying data distributions in wireless communication networks. VAEs was integrated with deep neural networks (DNNs) to enhance the ability of the DNN to adapt to complex and varying channel conditions. The adaptation resulted in a significant boost in the detection accuracy, which led to

improved performance in bit error rate (BER) and a signal-to-noise ratio gain of nearly 1 dB compared to traditional methods. The VAE is pivotal in overcoming the tasks associated with complex distribution characteristics and noise in industrial data. It was used to reconstruct process variables and incorporate Gaussian distribution constraints on the latent features. This effectively mitigated the impact of the complex data distribution. Also, by harnessing the potential of probabilistic distributions instead of traditional point estimates, the VAE enhanced the robustness of the soft sensor model. This assist to improve its ability to handle uncertainties and outliers (Wang & Wang, 2019; Boulila (2024) introduced a novel few-shot learning (FSL) technique for plant disease detection using a performer-attention mechanism and SwinTransformer. The FSL framework combines Performer attention and SwinTransformer architecture for efficient medical image classification using limited labeled data. The technique focused on key patches and masked regions prediction, the framework calculates similarity scores between new and known disease instances to enhance accurate classification with limited labeled examples. The technique was trained on the PlantVillage dataset, and it produced an accuracy of 99.12 % in a 10-shot learning scenario. To accurately classify complex multi-class diseases, Liu et al. (2024) introduced a patch-based neural network called PMLPNet for multi-class disease classification. PMLPNet captures spatial and channel contextual semantic features through carefully designed token and channel-mixing MLPs. The study utilized a dataset comprising 4510 images representing forty types of plant diseases across four different crops. The experimental results reveal that PMLPNet outperforms existing CNN models as it achieved a notable accuracy of 92.73 %. These studies demonstrate the versatility of VAEs in facilitating robust and efficient plant disease classification across different scenarios. Also, it highlights the potential of ViTs to complement feature extraction approaches to maintain robustness against image distortions and variations in real-world agricultural environments.

Building on these advancements, this study introduces a hybrid model that combines a Variational Autoencoder (VAE) for feature extraction with a Vision Transformer (ViT) for disease classification. The VAE reduces the dimensionality of high-resolution leaf images while preserving critical disease-related features, which helps to facilitate efficient processing. Subsequently, the ViT classifies these features by utilizing its global attention mechanism to enhance accuracy and robustness in disease detection. Therefore, by drawing insights from prior studies and incorporating innovative experimental techniques, this work emphasizes the potential of hybrid VAE-ViT models in advancing sustainable agriculture. It demonstrates their effectiveness in managing plant disease classification in diverse and challenging farming contexts which helps to contribute to improved agricultural practices and food security.

The following section prepares the ground for these two models with the goal of understanding their features and the crucial role they can play in improving the accuracy and robustness of plant disease identification.

2.1. Fundamental concepts of VAEs and VITs

2.1.1. Variational autoencoder (VAE)

VAE is an unsupervised deep learning model used for feature extraction, denoising, and dimensionality reduction. The VAE has two core components: an encoder and a decoder (Wei et al., 2020), as seen in Fig. 1. The encoder maps the input images (for example, plant leaves) to a lower-dimensional latent space, which can efficiently compress data into a relevant sample. The decoder reconstructs the images from this latent space (Han et al., 2019). When VAE is trained on a dataset (such as the leaves of plant images), it can learn to capture essential features and patterns present in the dataset (plant leaf images). Wang and Wang (2019) highlighted the use of VAE on the 450 K DNA methylation data of TCGA-LUAD and TCGA-LUSC to learn the latent representations of DNA methylation, and through their experiment, they showed that VAEs were very efficient in capturing the differential methylation patterns in subtypes of lung cancer. VAE was applied by Portillo et al. (2020) to reduce the dimension of a sample of spectra from the Sloan Digital Sky Survey (SDSS). They reported that VAE was able to capture the nonlinear relations between the latent parameters and the data better than principal component analysis (PCA). Similarly, other researchers have used VAE for other purposes such as data augmentation (Saldanha et al., 2022; Islam et al., 2021) and to denoise in diverse domains, such as image processing, audio, and text. VAEs present a significant advantage in this context due to their ability to independently extract meaningful representations from the data. This ability eliminates the necessity for manual feature engineering or different denoising procedures. Also, this flexibility makes VAEs a valuable technique in deep learning applications, especially plant disease identification. Essentially, the VAE's learn a compressed latent representation (z) of high-dimensional input data (plant leaf images). This helps to reduce the dimensionality of images while retaining disease-related features. The VAE is trained to maximize the Evidence Lower Bound (ELBO) of the log-likelihood of the data as depicted in (1).

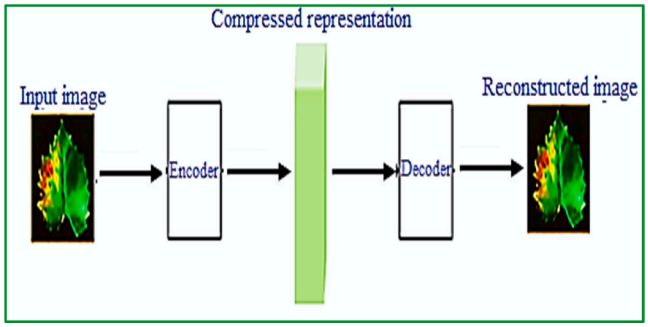


Fig. 1. The architecture of VAE.

$$\log p_\theta(x) \geq \mathbb{E}_{q_\varnothing(z|x)}[\log p_\theta(x|z)] - KL[q_\varnothing(z|x) \parallel p(z)], \tag{1}$$

where

$q_\varnothing(z|x)$ is the approximate posterior, parameterized by $\varnothing$,
$p_\theta(x|z)$ is the likelihood of data given the latent variables, parameterized by θ
$p(z)$ is the prior distribution, typically $N(0, I)$
$KL[. \parallel .]$ is the Kullback-Leibler divergence

The loss function for the VAE combines reconstruction loss and KL-divergence (which guarantees regularization in latent space).

$$\mathbf{L}\text{VAE} = \mathbb{E}_{q_\varnothing(z|x)}\left[\parallel x - \hat{x}\parallel^2\right] + \beta.\ KL[q_\varnothing(z|x) \parallel p(z)]. \tag{2}$$

In this study, the latent space representation $\boldsymbol{z}$ extracted by the VAE is used to reduce the dimensionality of high-resolution plant leaf images while preserving key features relevant to disease classification.

2.1.2. Vision transformers (VITs)

The ViT model typically utilizes transformer-like architecture over patches of image for image classification. ViT uses a self-attention mechanism to model the relationships between several image patches, which allows it to capture global relationships between image patches (Fig. 2). This potential is very effective for plant leaf disease identification, where disease symptoms can be distributed across the entire leaf. During patch processing, ViT splits an original input image into distinct patches and processes each patch independently. This allows ViT to learn more robust features that are less sensitive to image rotation, scaling, and other transformations. Also, the multi-head self-attention mechanism of ViT enables it to attend to different parts of an input image concurrently to learn more comprehensive representations. This is important for plant disease identification, where diverse parts of the

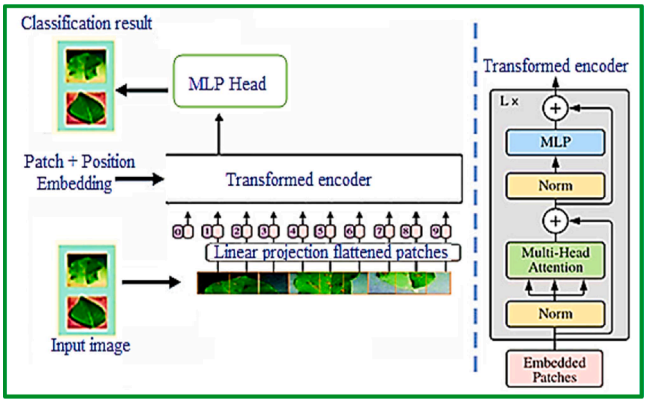


Fig. 2. The architecture of ViT (Isinkaye et al., 2024).

leaf could show varied symptoms. ViT is a lightweight and efficient design that is suitable for deployment on edge devices such as smartphones and embedded systems. It is therefore useful for plant disease identification, where the model is expected to process images in real-time to provide correct predictions. The ability of ViT to capture global relationships between image patches and its lightweight architecture present it as a promising approach for plant disease identification. Therefore, integrating ViT into this context is a novel step towards utilizing advanced methods to enhance automated plant disease diagnosis and management. Also, in recent times, many researchers have studied ViT models in different computer vision tasks due to their excellent performance (Alshammari et al., 2022; Thakur et al., 2023; Yang et al., 2023; Fan et al., 2023). Formally, the Vision Transformer (ViT) uses self-attention mechanisms to process image data. The image is divided into N patches, each flattened and embedded into a vector as expresses (3).

$$x_1 = Linear(Patch_i) + PositionalEnocoding_i \tag{3}$$

where Patch_i is the ith-image patch, and $\text{PositionalEncoding}_i$ adds positional information. The multi-head self-attention mechanism calculates attention scores to capture global dependencies as depicted in (4).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

where

Q, K, V are the query, key, and value matrices, d_k is the dimension of the keys.

The output of the self-attention mechanism is passed through a feed-forward network and layer normalization to form the final classification logits.

3. Materials and methods

In this study, we utilize a hybrid model that combines Variational Autoencoders (VAE) and Vision Transformers (ViT) to address the challenge of plant disease classification from leaf images. The primary aim is to harness the strengths of both models. VAE streamlines the high-dimensional data without losing crucial disease-related information, and ViT enhances the classification accuracy by effectively handling the complex inter-class variability in plant diseases through its global self-attention mechanism. Our approach provides a promising solution to the challenges posed by plant disease detection, which is vital for advancing precision agriculture and sustainable farming practices. The next section describes the dataset, preprocessing steps, and the architecture of the hybrid VAE-ViT model used in this study.

3.1. Dataset description

The proposed model is expected to identify the diseases of three plants: corn, potato, and tomato. The leaf images of these plants were obtained from the New Plant Diseases Dataset available at https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset. The dataset consists of 87,000 images of healthy and diseased crop leaves grouped into 38 different classes. Some samples of healthy and diseased leaf images of the corn, potato and tomato classes are shown in Fig. 3.

Majorly, the leaf images of corn, potato, and tomato plants were extracted from the New Plant Diseases Dataset on Kaggle. The extracted dataset contains a total of 27,872 leaf images. To address the issue of class imbalance, some classes of tomato images were excluded to create a more balanced dataset. We focused on five classes of tomato leaf images, this resulted in a total of twelve plant leaf disease classes used in the study. The dataset was divided into three subsets: 70 % for training, 15 % for validation, and 15 % for testing, as described in Table 1 and Fig. 4. The dimensions of the input images are specified as 224 × 224 × 3, which represent the height, width, and channel width, respectively.

3.2. Dataset preprocessing

To enhance the robustness and generalization of the proposed model, an on-the-fly data augmentation technique was meticulously implemented during each training epoch. Precisely, RandomFlip was utilized to introduce horizontal flips, RandomRotation applied random rotations of up to 10 %, and RandomZoom adjusted the images with up to 10 % zoom. The dynamic augmentation pipeline substantially enriched the variability of the input data as it allows the model to process diverse transformations in real time. These enhancements not only improved the ability of the model to generalize across unseen data but also emphasized its superior adaptability and performance in handling complex plant disease classification tasks.

3.3. Description of the proposed model

The hybrid model begins with plant leaf images serving as input. The dataset is preprocessed before they are processed by a Variational Autoencoder (VAE), which extracts latent features by reducing the dimensionality while retaining essential disease-related details. The extracted features are then fed into a Vision Transformer (ViT), which employs its self-attention mechanism to evaluate these features and classify the leaves as either healthy or diseased. This flow combines the strength of VAE in data compression with the robust classification capability of ViT. This guarantees high accuracy in plant disease detection. The flow diagram of the proposed model is shown in Fig. 5.

Fig. 3. Plant image samples from the New Plant Disease Dataset.

Table 1
Summary of the number of plant leaf images in the train, validation and test sets.

S/N	Plant leaf name	Plant leaf disease class	No. of images in train set	No. of images in validation set	No. of images in test set
1.	Corn	Cercospora leaf spot	14,363	308	308
		Common rust	1669	358	358
		Healthy corn leaves	1627	349	349
		Northern leaf blight	1670	342	342
2.	Potato	Potato Early blight	1697	364	364
		Potato Late blight	1697	364	364
		Healthy potato leaves	1596	342	342
3.	Tomato	Bacterial spot	1489	319	319
		Tomato Early blight	1680	360	360
		Healthy tomato leaves	1685	361	361
		Tomato Late blight	1620	347	347
		Leaf mold	1646	353	353
		Total	19,510	4181	4181
		Overall Total	27,872		

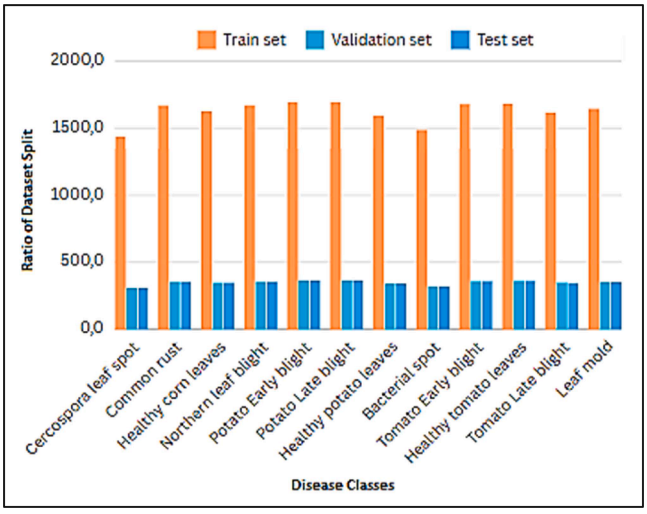


Fig. 4. Distributions of the number of images in the train, validation and test sets.

The feature extraction process entails that the encoder of the VAE maps the input plant image x to a compressed representation in a lower-dimensional latent space z. represented as:

$$z = \mu + \sigma \odot \varepsilon, \varepsilon \sim N(0, I) \tag{5}$$

where μ and σ are the learned mean and variance. ϵ is a random noise sampled from a normal distribution N(0, I) which introduces stochasticity to the representation. ⊙ is the element-wise multiplication. z is the latent vector used as the compressed feature representation of the image

In the classification phase, the latent representation z (the output of VAE encoder) is transformed into patch embeddings for the ViT. These embeddings are passed through the ViT for classification expressed as:

$$\hat{y} = ViT(z) \tag{6}$$

where $\hat{y}$ represents the predicted class output by the ViT, which classifies the plant leaf as either healthy or diseased based on the extracted features. The hybrid model is trained using a combined loss function that simultaneously optimizes feature extraction and classification accuracy.

For the training objective, VAE Reconstruction Loss (L_{VAE}) ensures that the VAE reconstructs the input image x accurately from the latent space z. This loss balances the fidelity of the extracted features. Classification Loss ($L_{classification}$) uses cross-entropy loss to ensure that the ViT correctly predicts the disease class labels. The overall loss function is expressed as:

$$L = L_{VAE} + a \cdot L_{classification} \tag{7}$$

3.3.1. *The design of the models*

This section provides a detailed breakdown of the model design process, it starts with explanations of the Variational Autoencoder (VAE) and Vision Transformer (ViT) as standalone components. The VAE is introduced as a tool for dimensionality reduction and latent feature extraction from high-resolution plant leaf images. The ViT is then explained as a classifier that uses its global attention mechanism to achieve robust and precise plant disease identification. After establishing the foundational understanding of each model, the section concludes with the integration of these two techniques into the proposed VAE-ViT hybrid model.

3.3.1.1. The design of the VAE. A Variational Autoencoder (VAE) was designed to reduce the dimensionality and extract features from plant leaf images as indicated in Fig. 6. The encoder comprises two convolutional layers with 32 and 64 filters (kernel size: 3, strides: 2, ReLU activation, and 'same' padding) to reduce spatial dimensions while capturing image features. These are followed by a dense layer (128 units, ReLU activation) for additional feature extraction and two separate dense layers generating the latent space's mean (z_mean) and log

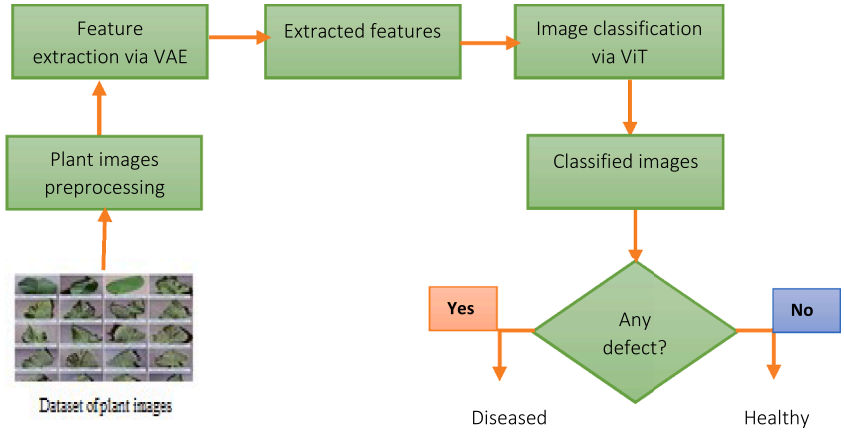


Fig. 5. The flow diagram of the hybrid VAE-ViT plant disease model.

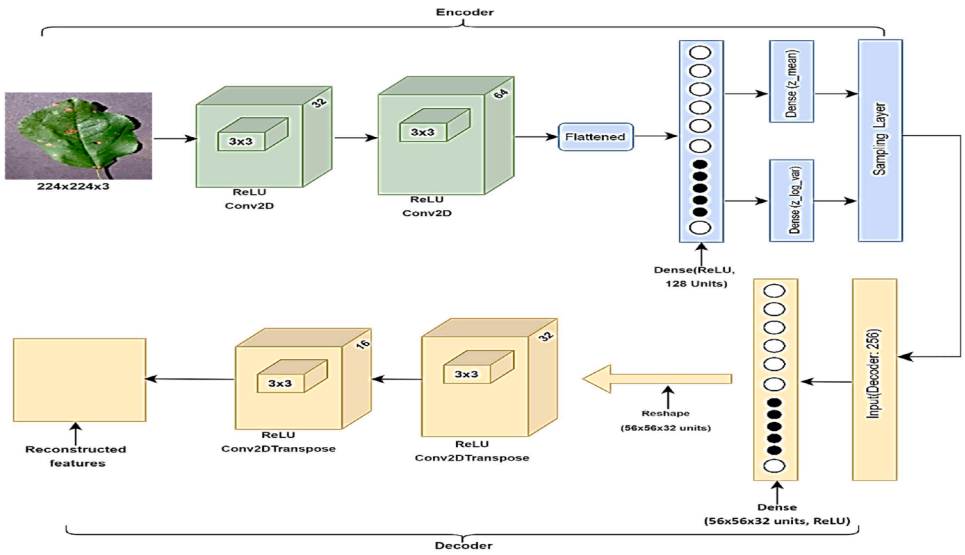


Fig. 6. The VAE architecture of the proposed model.

variance (z_log_var). The decoder reconstructs the original image through a fully connected layer that reshapes the latent vector into a 56 × 56 × 32 tensor. Three transposed convolutional layers (filters: 32 and 16; strides: 2; ReLU activation) and a final transposed convolutional layer (sigmoid activation) restore the image while normalizing the output. To determine optimal VAE parameters, sensitivity analysis was performed on latent dimensions, batch size, and learning rates. Results indicate that increasing latent dimensions beyond 256 showed diminishing returns.

3.3.1.2. The design of the ViT. The Vision Transformer (ViT) model incorporates an input layer, a pre-trained MobileNetV2 base, a global average pooling layer, a fully connected layer, a dropout layer (50 % dropout rate), and an output layer, as shown in Fig. 7. The MobileNetV2 serves as the feature extractor, it utilizes its pre-trained weights, which remain frozen during training to retain its learned features. Input images are passed to this base model, and the global average pooling layer reduces the spatial dimensions of the feature maps to a manageable size. A dropout layer is included to mitigate overfitting by randomly deactivating 50 % of neurons during training. The output layer employs a SoftMax activation function to produce a probability distribution across the classes to enhance disease classification. Sensitivity analysis revealed that a dropout rate of 50 % provided the best balance between performance and generalization.

3.3.1.3. The design of the hybrid VAE-ViT model. As shown in Fig. 8, the hybrid model starts by utilizing a dense layer that transforms the latent vector $\boldsymbol{z}$ into a higher-dimensional representation with dimensions 56 × 56 × 356. This is followed by a reshape layer that organizes this representation into a 56 × 56 × 356 tensor, that is suitable for further processing. Next, an upsampling module employs two Conv2DTranspose layers and one convolutional layer. The Conv2DTranspose layers progressively upsample the tensor using strides of 2 and ReLU activation to enhance the spatial resolution. The final convolutional layer ensures the output tensor has three channels that aligns with the RGB format of the original input images. Also, a resizing layer then adjusts the upsampled tensor to match the input size required by the Vision Transformer (ViT) model. The processed tensor is subsequently passed to the ViT, which analyzes the data and outputs class probabilities, which determines whether the plant leaf is healthy or diseased.

The model utilizes the Adam optimizer with an exponential decay learning rate schedule, initially set to 1e-4. The decay steps are set to 100,000 with a decay rate of 0.96 to ensure gradual and controlled adjustments to the learning rate during training. Categorical cross-entropy loss is used as the loss function, while classification accuracy serves as the primary evaluation metric. To enhance training efficiency and prevent overfitting, early stopping and model checkpointing callbacks were implemented. Table 2 outlines the training parameters for the proposed model in detail. Experiments on latent dimensions and

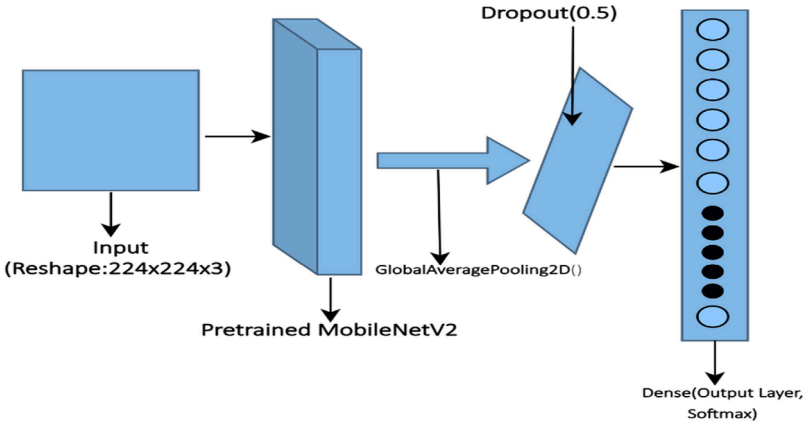


Fig. 7. The architecture of the proposed ViT model.

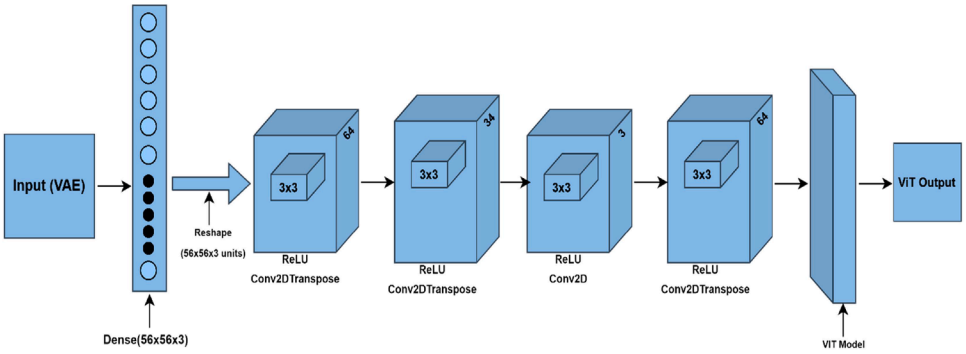


Fig. 8. The architecture of the proposed hybrid VAE-ViT model.

Table 2
Training parameters.

Parameter	Value
Epochs	15
Latent dimensions	256
Batch size	32
Number of classes	12
initial learning rate	1e-4
Decay steps	100,000
Decay rate	0.96

learning rates demonstrated that the chosen settings provided optimal trade-offs between accuracy and computational cost.

3.3.2. Complexity and uncertainty of the hybrid model

This section provides an analysis of the complexity and uncertainty of the proposed model. It highlights the computational requirements and robustness to parameter variations.

3.3.2.1. Computational complexity. The proposed hybrid model harnesses the computational efficiency of the VAE for dimensionality reduction and the ViT for classification. The ViT's global attention mechanism contributes a time complexity of $O(n^2)$, where n is the number of image patches, this makes it computationally intensive for high-resolution images. The memory complexity is influenced by the size of the latent space in the VAE and the number of parameters in the ViT. Overall, the training and inference processes of the model were optimized to ensure scalability and practicality for real-world applications.

All experiments were conducted on a high-performance cluster computer equipped with the following specifications:

- Processors: 2 × Intel Xeon E5–2697A v4
- Memory: 512 GB of 2.4 GHz DDR4 RAM

This computational infrastructure facilitated the training and evaluation of the hybrid model, it ensures the timely completion of experiments despite the model's computational demands. The computational requirements for training the hybrid model include significant memory resources to accommodate the ViT's global attention mechanism and the VAE's dimensionality reduction tasks. During training, the model utilized a high percentage of the available RAM and sustained moderate CPU utilization. The training duration for the hybrid model, across 15 epochs with a batch size of 32 and 12 output classes, was 4.12 h for the dataset used in this study. These requirements underline the importance of access to robust computational resources for implementing the proposed model effectively.

3.3.2.2. Model parameter uncertainty. Parameter uncertainty was addressed through careful selection and evaluation of hyperparameters, such as learning rates, latent dimensions, and batch sizes. Sensitivity analyses were performed to assess the impact of parameter variations on model performance. Key findings include:

- Latent space dimensions directly influence the balance between feature richness and computational load.
- Learning rate schedules significantly affect convergence stability and training duration.

3.4. Evaluation of the hybrid VAE-ViT model

This study employs a range of evaluation metrics to assess the model's performance. These include classification accuracy, recall (sensitivity), precision, and F1-score (Isinkaye et al., 2015) which are described as follows.

- Classification accuracy measures how well the model differentiates between positive and negative classes as depicted in Eq. (5). A high accuracy score implies that the model makes a substantial number of correct predictions, while a low accuracy score shows that the model makes a lot of incorrect predictions.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8}$$

Where:

TP (True Positives): Correctly predicted positive instances.
TN (True Negatives): Correctly predicted negative instances.
FP (False Positives): Incorrectly predicted positive instances.
FN (False Negatives): Incorrectly predicted negative instances.

- Precision measures the accuracy of positive predictions of the model and it is computed as expressed in Eq. (6):

$$\text{Precision} = \frac{(TPs)}{(TPs + FPs)} \tag{9}$$

- Sensitivity measures the model's ability to detect all true positive instances (diseased samples), it is computed as seen in Eq. (7):

$$Sensitivity = \frac{(TPs)}{(TPs + FNs)} \tag{10}$$

- F1 score is a balanced measure of precision and recall, calculated as Eq. (8):

$$F1 = 2 \times \frac{(Precision \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (11)$$

4. Results and discussion

This section presents the results of our study on the effectiveness of the hybrid VAE-ViT model for multi-class plant disease classification where VAEs and ViTs were integrated. The proposed hybrid model was trained on a dataset containing images from three crops: corn, potato, and tomato. Corn leaf images are classified into four disease categories, potato images into three, and tomato images into five, as described in Table 2.

4.1. Performance evaluation of the hybrid model across all plant diseases

Fig. 9 illustrates the training and validation accuracy and loss of the proposed model. The results demonstrate that the model generalizes effectively to unseen images, as evidenced by the minimal gap between the training and validation accuracy and loss. This consistency indicates that the model avoids significant overfitting while maintaining strong predictive performance. It also reflects the ability of the VAE to retain essential image features during dimensionality reduction and the ViT's capability to learn complex patterns and dependencies. These outcomes validate the reliability and robustness of the proposed system for plant disease classification. Thus, it aligns with the objective of the study to develop an accurate and reliable classification framework.

The results in Table 3 show the effectiveness of the hybrid model in multiclass plant disease classification. With a 93.2 % classification accuracy, the model was able to handle binary classifications and excelled in identifying multiple disease classes across varied conditions, such as changes in lighting and background. This performance outperforms traditional CNNs, especially when faced with environmental disturbances such as changes in lighting or background. The VAE reduces image dimensionality while preserving essential features, the ViT captures complex patterns and long-range dependencies. Thus, the hybrid model offers improved scalability, with the ability to handle large datasets more efficiently. It achieved greater classification precision in complex, multi-class plant disease scenarios. This makes the hybrid model particularly effective in real-world agricultural applications, where numerous diseases across different plants need to be identified quickly and accurately.

Also, the high precision of the model (93.8 %) is very critical for real-world agricultural applications, it minimizes false positives and ensures that healthy plants are not misclassified as diseased. The reduced false positive rate lowers the risk of unnecessary interventions, which makes the model a reliable tool for practical deployment. These findings validate the robustness of the hybrid VAE-ViT framework as an efficient and accurate solution for plant disease classification. However, the slightly lower sensitivity of the model for certain diseases warrants further investigation. This may be due to subtle visual differences between healthy and diseased samples or an imbalance in the dataset.

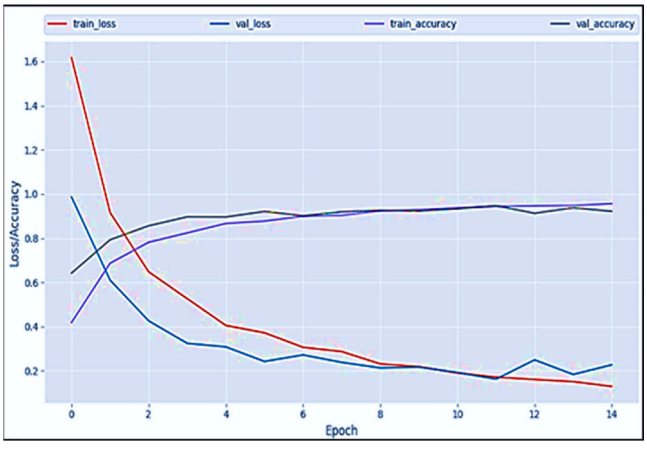


Fig. 9. Training and validation accuracy and loss of the proposed model.

Table 3
Performance of the hybrid VAE-ViT model.

Crop name	Test accuracy %	Sensitivity %	Precision %	F1 Score %
Corn, Potato &Tomato	93.2	93.2	93.8	93.1

4.2. Performance of the hybrid model on each plant disease class

The results produced by the proposed model for each disease class are presented in Table 4 and Fig. 10, respectively. As shown, the test accuracy produced for each disease is remarkable. The model produced the best accuracies of 99.6 % and 99.3 % for both Common rust and Healthy tomato leaves, respectively. The model achieved an accuracy ranging from 94.2 % to 98.3 % for the other diseases, as seen in the table. The accuracy shows that the model correctly classified between 94 and 99 samples out of 100 for each disease. These results underscore the reliability of the model in detecting diverse disease patterns with minimal errors. Generally, developing a model with high accuracy is critical in crop disease classification. Models with low classification accuracy can lead to misclassifications and, consequently, incorrect treatments and reduced production yields. The model's high accuracy suggests it can conveniently identify crop diseases as demonstrated in its excellent performance across all disease classes.

The precision of the proposed model is also notable. Cercospora leaf spot and Healthy tomato leaves achieve the highest precision at 98.7 % and 97.6 %, respectively. Precision for other disease classes ranges from 91.9 % to 95.9 %. This means that out of all samples predicted as diseased, 91.9 % to 98.7 % were correctly identified. High precision ensures that most positive predictions are true positives which reduces unnecessary treatments and associated costs as well as environmental impacts.

The result of sensitivity, or recall, is equally striking, as cercospora leaf spot and healthy tomato leaves achieved the highest sensitivity of 99.7 % and 100.0 %, respectively. The model produced a sensitivity ranging from 93.7 % to 98.9 %. for other disease classes. This sensitivity shows that the model can detect between 93.7 % and 100 % of actual diseased images, which demonstrate the capability of the model to detect actual diseased instances effectively, this minimizes the likelihood of undetected infections.

Table 4
Performance of the proposed model for each plant disease class.

Name of disease	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
Cercospora leaf spot	98.3	98.7	99.7	99.2
Common rust	99.3	98.6	99.7	99.2
Healthy corn leaves	99.6	98.6	100.0 %	99.3
Northern leaf blight	93.3	77.7	98.2	87.1
Potato Early blight	98.3	95.9	98.9	97.4
Potato Late blight	96.3	94.1	98.5	96.3
Healthy potato leaves	94.2	92.3	93.7	92.9
Bacterial spot	96.3	94.5	97.7	96.1
Tomato Early blight	94.7	92.8	95.5	94.1
Healthy tomato leaves	99.3	97.6	100.0 %	98.8
Tomato Late blight	95.7	91.9	97.7	94.7
Leaf mold	96.3	97.9	97.6	97.6
Average	**96.8**	**94.2**	**98.1**	**96.1**

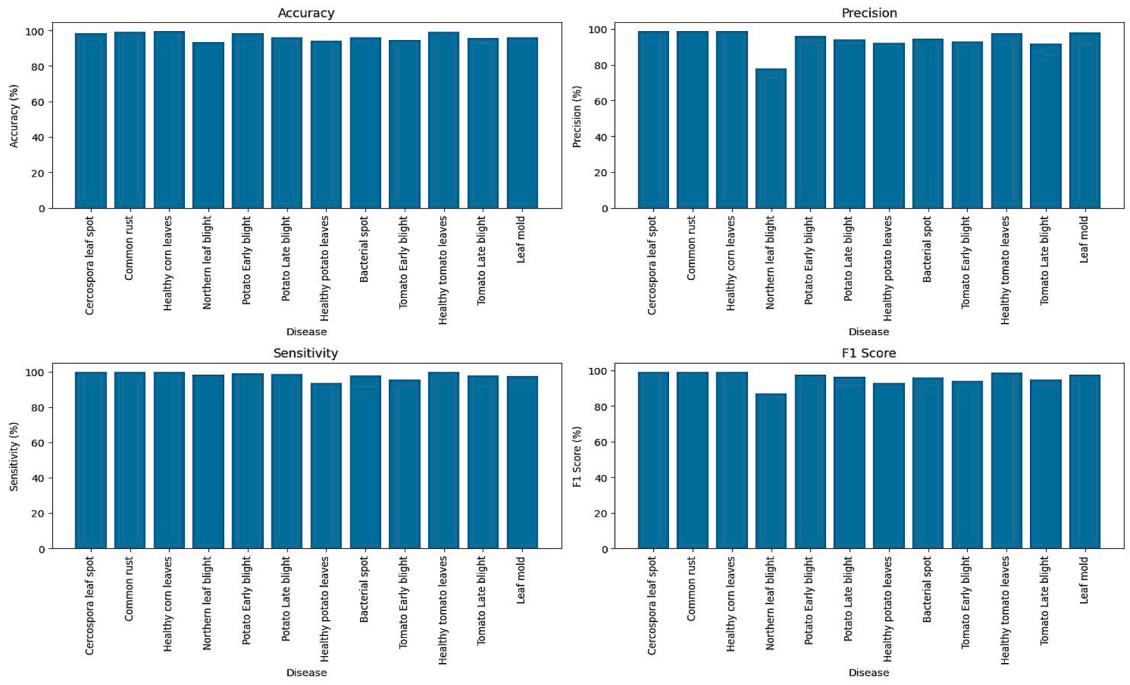


Fig. 10. Performance of the proposed model for different plant disease classes.

The F1 score produced by the model is also satisfactory. As shown in the results, the model achieved the best F1 score of 99.3 % for healthy corn leaves. It produced an F1 score ranging from 92.9 % to 97.4 % for the other diseases. This highlights the reliability of the model in managing the trade-off between false positives and false negatives.

In comparing the performance of the proposed model across the 12 plant disease classes, it was observed that Cercospora leaf spot and healthy tomato leaves consistently demonstrate the highest accuracy, precision, sensitivity, and F1 score. The slight performance dip for Northern leaf blight (93.3 % accuracy and 87.1 % F1 score) indicates potential challenges in distinguishing this class. This could stem from overlapping visual features between Northern leaf blight and other disease categories, necessitating further refinement in feature extraction and class separation. Integrating auxiliary models might enhance performance for this class. Overall, the high performance produced by the proposed model across all the diseases shows that it is an effective and reliable tool for the detection of plant leaf diseases.

4.2.1. *Confusion matrix results*

The confusion matrix for the hybrid model is shown in Fig. 11. It reveals the ability of the VAE-ViT model to classify plant diseases across 12 classes with varying levels of accuracy. Each class is represented in a row, with its corresponding predictions in columns. A closer analysis shows the patterns of strengths and challenges, which provides reasons for the observed performance.

Cercospora leaf spot and common rust: These two classes exhibit remarkable classification accuracy, with 304 and 355 correctly predicted instances, respectively. Misclassifications are minimal. This

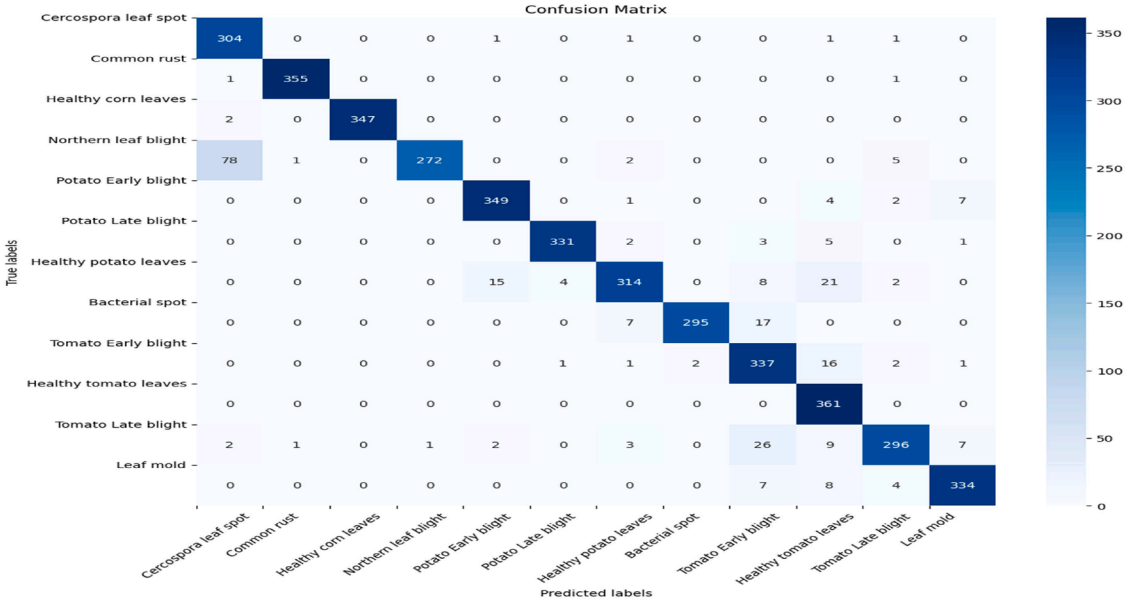


Fig. 11. Confusion matrix for the hybrid model.

suggests that the model successfully captures the distinct visual features associated with these diseases. The combination of VAEs and ViTs likely contributes to this success. VAEs reduce the complexity of input data while retaining critical disease features, and ViTs identified spatial relationships and global patterns.

Healthy corn leaves and healthy tomato leaves: These classes also show excellent classification performance, with 347 and 361 correct predictions, respectively. The near-perfect accuracy for healthy tomato leaves highlights the ability of the model to differentiate healthy leaves from diseased conditions. This result underscores the importance of distinct visual characteristics, which the hybrid model can accurately interpret.

Northern leaf blight and tomato late blight: These classes demonstrate the challenge of the model in handling visually similar diseases. Northern leaf blight, for instance, is often misclassified as Cercospora leaf spot or Tomato Late blight. This is likely due to overlapping visual features between these conditions, which the model struggles to disentangle fully. Similarly, Tomato Late blight's confusion with Northern leaf blight reflects the need for improved feature extraction and enhanced data diversity.

Potato early blight and potato late blight: Both blight classes exhibit high accuracy, with minor misclassifications spread across related classes. These results indicate the capacity of the model to identify distinctive features but also highlight its sensitivity to class differences and subtle variations in disease presentations.

Bacterial spot, tomato early blight, and healthy potato leaves: Although these classes exhibit strong overall performance, some confusion arises between Bacterial Spot, Tomato Early Blight, and other leaf diseases. Notably, the model occasionally misclassifies Bacterial Spot as Tomato Early Blight. This shows that the two diseases share visual similarities that the model could further distinguish. Also, the model experiences some confusion between healthy potato leaves and Potato Early Blight/Tomato Early Blight, suggesting that the boundaries between these classes may be subtle and require further refinement.

4.3. Insights and implications

The results of this study underscore the effectiveness and adaptability of the hybrid VAE-ViT model in addressing the complex challenges of plant disease detection across diverse agricultural scenarios. The model demonstrates exceptional classification performance for most disease classes, owing to the synergistic interplay between Variational Autoencoders (VAEs) and Vision Transformers (ViTs).

4.3.1. Strengths of the model

The hybrid VAE-ViT model shows exceptional performance in detecting plant disease classes with distinct visual features, such as Common rust, Healthy corn leaves, and Healthy tomato leaves. This success stems from the complementary roles of VAEs and ViTs, which work together to enhance the robustness and accuracy of the model. VAEs play a pivotal role by reducing the dimensionality of input data while retaining essential disease-related features, which ensures the model focuses on relevant information and remains resilient to overfitting. The dimensionality reduction simplifies the learning task and makes it more efficient and effective. Meanwhile, ViTs utilize their global attention mechanisms to identify and analyze complex spatial relationships and dependencies across entire images. This capability enables the model to understand subtle disease patterns, even in high-dimensional datasets.

The hybrid architecture's ability to surpass traditional CNN-based methods is particularly noteworthy. CNNs often struggle with capturing global dependencies and handling the complexities of large datasets. In contrast, the VAE-ViT model excels by combining dimensionality reduction with global attention, resulting in high levels of accuracy, precision, and sensitivity. These strengths make the model a reliable and scalable tool for multi-class plant disease diagnosis, demonstrating its potential as a cutting-edge solution for practical agricultural applications.

4.3.2. Practical applications and scalability

The robust performance of the hybrid model ensures its practical applicability in real-world agricultural settings. Farmers can harness its accurate and timely disease detection capabilities to make informed decisions that can assist in reducing crop losses and enhancing productivity. The scalability of the model is evident in its ability to handle different disease classes. This makes it adaptable for deployment across different regions, crop types, and environmental conditions. The adaptability also positions the hybrid VAE-ViT model as a transformative tool for agricultural disease management.

4.3.3. Opportunities for continued advancement

While the overall performance of the model is strong, specific disease classes, such as Northern leaf blight and Tomato Late blight, exhibit minor misclassifications. These challenges are likely due to subtle interclass similarities and class imbalances in the training data. Addressing these issues presents an opportunity to further enhance the model's robustness. Future refinements could include:

- Class-specific data augmentation techniques to expand feature diversity.
- Advanced loss functions tailored for imbalanced datasets.
- Increased diversity in training datasets to capture a broader range of disease variations.

4.3.4. Technological and computational considerations

Despite its strong performance, the reliance on high-performance computational resources, such as GPUs or clusters, may limit accessibility for resource-constrained environments. However, this challenge can be addressed by optimizing the model for deployment on lower-resource systems or utilizing cloud-based solutions to democratise access.

4.3.5. Prospects

The hybrid VAE-ViT model represents a significant advancement in plant disease detection, demonstrating its potential as a scalable and efficient tool for agricultural disease management. The model's strengths far outweigh its limitations, as it consistently outperforms traditional methods in terms of accuracy, sensitivity, and practical applicability. Future enhancements, such as incorporating ensemble methods, expanding the model to include additional plant species, and evaluating its performance on diverse datasets, will solidify its role as a cutting-edge solution for modern agriculture.

In summary, the hybrid VAE-ViT model not only addresses current challenges in plant disease detection but also lays the foundation for future innovations in agricultural technology. Its demonstrated reliability and scalability underscore its transformative potential in achieving sustainable crop health management.

4.4. Comparison of the proposed model with previous studies

This section provides a comprehensive comparison of the proposed hybrid model against five previous studies, as summarized in Table 5. Barman et al. (2024) developed a model using a single plant species, achieving accuracy, precision, sensitivity, and F1 scores of 91.0 %, 91.0 %, 89.0 %, and 91.0 %, respectively. While their results are commendable for a single-plant context, the proposed hybrid model significantly outperforms Barman et al.'s model across all metrics, demonstrating superior capability in handling a broader range of plant diseases. Hassan and Maji (2022) focused on cassava and reported an accuracy of 76.7 %. This underscores a substantial improvement with our proposed model, which surpasses their performance and shows better generalization across multiple plant species and disease conditions.

Table 5
Comparison of existing models with the proposed model.

Author	Technique	No of Plant	Dataset source	Accuracy %	Precision %	Sensitivity %	F1 Score %
Barman et al. (2024)	ViT	1 -Tomato	Plant village dataset	91.0	91.0	89.0	91.0
Hassan and Maji (2022)	CNN (Inception and ResNet)	1-Cassava	Plant village dataset	76.6	–	–	–
Ahmad et al. (2020)	CNN	1- Tomato	Plant village dataset	–	84.5	90.6	87.6
Shoaib et al. (2022)	CNN based on CANet architecture	3- Pepper, Potato and Tomato	Plant village dataset	93.0	91.8	91.8	91.8
Reddy et al. (2023)	CNN	38-plant species	Plant village dataset	90.0	–	–	–
Proposed model	VAE and ViT	3- Corn, Potato and Tomato	Plant village dataset	93.2	93.2	93.8	93.1

Ahmad et al. (2020) achieved a precision of 84.5 %, a sensitivity of 90.6 %, and an F1 score of 86.5 % with their single-plant model. Despite the high precision, their model's performance is notably lower compared to the proposed hybrid model. The hybrid models enhanced precision and F1 score reflects the effectiveness of combining Variational Autoencoders (VAE) and Vision Transformers (ViT) for more reliable plant disease classification. Shoaib et al. (2022) employed a model with three plants (pepper, potato, and tomato), obtaining accuracy, precision, sensitivity, and F1 scores of 93.0 %, 91.8 %, 91.8 %, and 91.8 %, respectively. The proposed model not only matches but exceeds these metrics, demonstrating the significant advantages of integrating VAE for feature extraction and ViT for classification, which enhances the model's classification accuracy and robustness.

Reddy et al. (2023) utilized a large-scale dataset with thirty-eight plants, achieving an accuracy of 90.0 %. In contrast, the proposed hybrid model, trained on a dataset of three plants with 12 classes and over 22,000 images, achieves superior performance metrics: an accuracy of 93.2 %, precision of 93.2 %, sensitivity of 93.8 %, and an F1 score of 93.1 %. This comparison highlights the exceptional robustness and scalability of the proposed model. In essence, the proposed hybrid model significantly outperformed existing models across multiple metrics. While Barman et al.(2024) achieved commendable results with a single plant species, our model excels by effectively handling a wider range of plant diseases while achieving an accuracy of 93.2 % compared to Reddy et al. with 90.0 % and a larger dataset. Furthermore, it beats the 76.7 % accuracy reported by Hassan and Maji (2022) for cassava, which reflects its superior generalization capabilities across diverse plant species and conditions. The result of Ahmad et al. (2020) with high precision of 84.5 % is outmatched by our model. This emphasizes the enhanced reliability afforded by the integration of Variational Autoencoders (VAE) and Vision Transformers (ViT). Furthermore, even though Shoaib et al. reported striking metrics with three plant species, our hybrid approach not only matches their performance but also offers a robust framework for improved classification accuracy and resilience in practical applications. Also, it will be a highly effective solution for practical plant disease classification, especially where farmers grow multiple crops simultaneously.

In summary, the study presents a hybrid architecture that integrates VAEs and ViTs for plant disease identification. It effectively addresses the limitations of traditional CNNs by improving computational efficiency and capturing long-range dependencies. The model demonstrates effectiveness in identifying multiple plant diseases simultaneously as it achieved a significant classification accuracy of 93.2 % across three crops (corn, potato, and tomato), substantially outperforming existing methods. Furthermore, the use of VAEs for dimensionality reduction enables the model to efficiently manage large agricultural datasets without sacrificing essential features. This makes it a scalable solution for real-world applications. Also, with the utilization of ViTs, the model also improves its ability to capture both global relationships and local details in images. This assists in enhancing robustness against noise and high-resolution inputs commonly encountered in agricultural contexts. Finally, the model's capacity for accurate multi-crop disease identification positions it as a valuable tool for farmers and agricultural stakeholders to accelerate timely disease identification and better crop management. This will also help to improve food security and economic benefits in agriculture.

5. Conclusion

This study introduced a DL-based approach to plant disease identification by combining the strengths of VAE for feature extraction and ViT for classification to overcome the limitations of CNN techniques. The model was trained on three classes of plant leaf images which are corn, tomato and potato. Experimental results confirm that the proposed hybrid model achieved a high accuracy of 93.2 % in correctly and efficiently identifying plant leaf diseases. Additionally, the robustness of the model was also validated with other metrics such as precision, sensitivity and F1-score respectively. The results emphasize that the performance of the hybrid model is balanced in accurately and efficiently identifying diseases in plant images with precision, 93.6 %; sensitivity, 93.2 %; and Fi-score, 93.1 %. The comprehensive analysis of each disease in the three classes of plant leaf images showed remarkable accuracy across all diseases. For example, Cercospora leaf spot attained a very high accuracy of 99.3 %.

This confirms that the hybrid model successfully harnessed the strengths of both VAEs and ViTs to enhance the robustness and accuracy of plant disease identification. It effectively manages high-dimensional data, extracts detailed features, and ensures precise classification across multiple disease classes even with noisy images. Therefore, the hybrid technique not only overcomes some of the limitations associated with CNNs but also provides a more reliable solution for detecting multiple diseases across different plants as seen in its superior performance in both accuracy and precision. The proposed hybrid model also significantly outperformed some existing models across multiple metrics.

The performance of the model also underscores its potential as a reliable tool for stakeholders, especially farmers. It can improve crop management, reduce economic losses and enhance food security. Also. The research prepares the foundation for future studies in smart agriculture which reinforces its contribution to both academic research and real-world application

In the future, we hope to explore the inclusion of other state-of-the-art deep learning approaches and increase the size of our dataset to accommodate more classes of plant leaf diseases. We hope to equally refine the model to further enhance its performance accuracy and applicability to ensure its continuous importance as a cutting-edge solution in plant disease identification.

CRediT authorship contribution statement

Folasade Olubusola Isinkaye: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Michael Olusoji Olusanya:** Conceptualization, Supervision, Writing – review & editing. **Ayobami Andronicus Akinyelu:** Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Folasade O. Isinkaye reports financial support was provided by National Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the South African National Research Foundation (NRF) under Grant Number PSTD2204113035, through the Innovation Postdoctoral Fellowship award and the hosting institution, Sol Plaatje University, Kimberley, Republic of South Africa.

Data availability

Data will be made available on request.

References

Ahmad, I., Hamid, M., Yousaf, S., Shah, S. T., & Ahmad, M. O. (2020). Optimizing pretrained convolutional neural networks for tomato leaf disease detection. *Complexity, 2020*(1), Article 8812019.

Ahmed, S. T., Barua, S., Fahim-Ul-Islam, M., & Chakrabarty, A. (2024). Enhancing precision in rice leaf disease detection: A transformer model approach with attention mapping. In *2024 International conference on advances in computing, communication, electrical, and smart systems (iCACCESS)* (p. 1).

Alshammari, H., Gasmi, K., Ltaifa, I. B., Krichen, M., Ammar, L. B., & Mahmood, M. A. (2022). Olive disease classification based on vision transformer and CNN models. *Computational Intelligence and Neuroscience*.

Anstine, D. M., & Isayev, O. (2023). Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society, 145*(16), 8736–8750.

Arsenovic, M., Karanovic, M., Sladojevic, S., Anderla, A., & Stefanovic, D. (2019). Solving current limitations of deep learning-based approaches for plant disease detection. *Symmetry, 11*(7), 939.

Barman, U., Sarma, P., Rahman, M., Deka, V., Lahkar, S., Sharma, V., & Saikia, M. J. (2024). ViT-SmartAgri: Vision transformer and smartphone-based plant disease detection for smart agriculture. *Agronomy, 14*(2), 327.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 12*(5), e1608.

Boulila, W. (2024). An approach based on performer-attention-guided few-shot learning model for plant disease classification. *Earth Science Informatics, 2024*, 1–13.

De Silva, M., & Brown, D. (2023). Multispectral plant disease detection with vision transformer–convolutional neural network hybrid approaches. *Sensors, 23*(20), 8531.

Elbattah, M., Loughnane, C., Guérin, J. L., Carette, R., Cilia, F., & Dequen, G. (2021). Variational autoencoder for image-based augmentation of eye-tracking data. *Journal of Imaging, 7*(5), 83.

Fan, R., Alipour, K., Bowd, C., Christopher, M., Brye, N., Proudfoot, J. A., & Zangwill, L. M. (2023). Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization. *Ophthalmology Science, 3*(1), Article 100233.

Faisal, M., Leu, J. S., & Darmawan, J. T. (2023). Model selection of hybrid feature fusion for coffee leaf disease classification. *IEEE Access, 12*, 62281–62291.

Guo, Y., Zhang, J., Yin, C., Hu, X., Zou, Y., Xue, Z., & Wang, W. (2020). Plant disease identification based on deep learning algorithm in smart farming. *Discrete Dynamics in Nature and Society, 2020*(1), Article 2479172.

Han, K., Wen, H., Shi, J., Lu, K. H., Zhang, Y., Fu, D., & Liu, Z. (2019). Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage, 198*, 125–136.

Harshvardhan, G. M., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review, 38*, Article 100285.

Hassan, S. M., & Maji, A. K. (2022). Plant disease identification using a novel convolutional neural network. *IEEE Access, 10*, 5390–5401.

Islam, Z., Abdel-Aty, M., Cai, Q., & Yuan, J. (2021). Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention, 151*, Article 105950.

Isinkaye, F. O., Olusanya, M. O., & Singh, P. K. (2024). Deep learning and content-based filtering techniques for improving plant disease identification and treatment recommendations: A comprehensive review. *Heliyon, 10*(9), e29583.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal, 16*(3), 261–273.

Jung, Y., Kim, Y., Choi, Y., & Kim, H. (2018). Joint learning using denoising variational autoencoders for voice activity detection. *Interspeech*, 1210–1214.

Kaur, P., Pannu, H. S., & Malhi, A. K. (2019). Plant disease recognition using fractional-order Zernike moments and SVM classifier. *Neural Computing and Applications, 31*, 8749–8768.

Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review, 56*(Suppl. 3), 2917–2970.

Kotwal, J., Kashyap, R., & Pathan, S. (2023). Agricultural plant diseases identification: From traditional approach to deep learning. *Materials Today: Proceedings, 80*, 344–356.

Li, Y., Mao, H., Girshick, R., & He, K. (2022). Exploring plain vision transformer backbones for object detection. In *European conference on computer vision* (pp. 280–296). Springer Nature Switzerland.

Liu, L., Chang, J., Qiao, S., Xie, J., Xu, X., & Qiao, H. (2024). PMLPNet: Classifying multi-class pests in wild environment via a novel convolutional neural network. *Agronomy, 14*(8), 1729.

Ma, X., Lin, Y., Nie, Z., & Ma, H. (2020). Structural damage identification based on unsupervised feature-extraction via Variational Auto-encoder. *Measurement, 160*, Article 107811.

Majikumna, K. U., Zineddine, M., & Alaoui, A. E. H. (2024). FLVAEGWO-CNN: Grey wolf optimisation-based CNN for classification of olive leaf disease via focal loss variational autoencoder. *Journal of Phytopathology, 172*(6), e13438.

Mahmud, M. S., Huang, J. Z., & Fu, X. (2020). Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. *International Journal of Computational Intelligence and Applications, 19*(1), Article 2050002.

Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., & Xue, H. (2022). Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12042–12051).

Mauricio, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences, 13*(9), 5521.

Muluneh, M. G. (2021). Impact of climate change on biodiversity and food security: A global perspective—A review article. *Agriculture & Food Security, 10*(1), 1–25.

Nagachandrika, B., Prasath, R., & Joe, I. P. (2024). An automatic classification framework for identifying types of plant leaf diseases using multi-scale feature fusion-based adaptive deep network. *Biomedical Signal Processing and Control, 95*, Article 106316.

Ngugi, L. C., Abelwahab, M., & Abo-Zahhad, M. (2021). Recent advances in image processing techniques for automated leaf pest and disease recognition—A review. *Information Processing in Agriculture, 8*(1), 27–51.

Omondi, G., & Olwal, T. O. (2023). Variational autoencoder-enhanced deep neural network-based detection for MIMO systems. *e-Prime-Advances in Electrical Engineering. Electronics and Energy, 6*, Article 100335.

Orchi, H., Sadik, M., & Khaldoun, M. (2021). On using artificial intelligence and the internet of things for crop disease detection: A contemporary survey. *Agriculture, 12*(1), 9.

Panigrahi, K. P., Das, H., Sahoo, A. K., & Moharana, S. C. (2020). Maize leaf disease detection and classification using machine learning algorithms. In *Progress in computing, analytics and networking: Proceedings of ICCAN 2019* (pp. 659–669). Springer.

Portillo, S. K., Parejko, J. K., Vergara, J. R., & Connolly, A. J. (2020). Dimensionality reduction of SDSS spectra with variational autoencoders. *Monthly Notices of the Royal Astronomical Society, 497*(3), 3161–3172.

Prakash, M., Krull, A., & Jug, F. (2020). Fully unsupervised diversity denoising with convolutional variational autoencoders. arXiv preprint arXiv:2006.06072.

Reddy, S. R., Varma, G. S., & Davuluri, R. L. (2023). Optimized convolutional neural network model for plant species identification from leaf images using computer vision. *International Journal of Speech Technology, 26*(1), 23–50.

Rizzo, D. M., Lichtveld, M., Mazet, J. A., Togami, E., & Miller, S. A. (2021). Plant health and its effects on food safety and security in a One Health framework: Four case studies. *One Health Outlook, 3*(1), 6.

Saldanha, J., Chakraborty, S., Patil, S., Kotecha, K., Kumar, S., & Nayyar, A. (2022). Data augmentation using variational autoencoders for improvement of respiratory disease classification. *PLoS ONE, 17*(8), Article e0266467.

Sajitha, P., Andrushia, A. D., Anand, N., & Naser, M. Z. (2024). A review on machine learning and deep learning image-based plant disease classification for industrial farming systems. *Journal of Industrial Information Integration*, Article 100572.

Shelar, N., Shinde, S., Sawant, S., Dhumal, S., & Fakir, K. (2022). Plant disease detection using CNN. In , *44*. *ITM web of conferences* (p. 03049). EDP Sciences.

Shoaib, M., Shah, B., Hussain, T., Ali, A., Ullah, A., Alenezi, F., et al. (2022). A deep learning-based model for plant lesion segmentation, subtype identification, and survival probability estimation. *Frontiers in Plant Science, 13*, Article 1095547.

Tao, Y., Tao, H., Zhuang, Z., Stojanovic, V., & Paszke, W. (2024). Quantized iterative learning control of communication-constrained systems with encoding and decoding mechanism. *Transactions of the Institute of Measurement and Control, 46*(10), 1943–1954.

Thakur, P. S., Chaturvedi, S., Khanna, P., Sheorey, T., & Ojha, A. (2023). Vision transformer meets convolutional neural network for plant disease classification. *Ecological Informatics, 77*, Article 102245.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., & Herath, D. (2023). Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence, 126*, Article 106669.

Wang, Z., & Wang, Y. (2019). Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. *BMC Bioinformatics, 20*(18), 1–7.

Wei, R., Garcia, C., El-Sayed, A., Peterson, V., & Mahmood, A. (2020). Variations in variational autoencoders—A comparative evaluation. *IEEE Access, 8*, 153651–153670.

Xia, Y., Chen, C., Shu, M., & Liu, R. (2023). A denoising method of ECG signal based on variational autoencoder and masked convolution. *Journal of Electrocardiology, 80*, 81–90.

Yang, Y., Zhang, L., Ren, L., & Wang, X. (2023). MMViT-Seg: A lightweight transformer and CNN fusion network for COVID-19 segmentation. *Computers in Biology and Medicine, 230*, Article 107348.

Yao, R., Liu, C., Zhang, L., & Peng, P. (2019). Unsupervised anomaly detection using variational auto-encoder based feature extraction. In *2019 IEEE international conference on prognostics and health management (ICPHM)* (pp. 1–7).

Yousuf, A., & Khan, U. (2021). Ensemble classifier for plant disease detection. *International Journal of Computer Science and Mobile Computing, 10*(1), 14–22.