GMR
The Original

TRANSWALKCOTNET: A SWIN TRANSFORMER AND DEEPWALK-BASED CROSS-ATTENTION FUSION FRAMEWORK FOR COTTON PLANT DISEASE IDENTIFICATION HYBRID GRAPH + TRANSFORMER MODEL

Hina Shafi[1]*, Ali Ghulam[1], Mir. Sajjad Hussain Talpur[1], Hira Sajjad[1], Zulfikar Ahmed Maher[1], Dr. Riaz Ali Buriro[2], Rahu Sikander[3]

1. Information Technology Centre, Sindh Agriculture University, Sindh, Pakistan
2. Department of Statistics, Sindh Agriculture University, Sindh, Pakistan
3. Department of Computer Science & Software Engineering Jinnah University for Women, Sindh, Pakistan

Correspondence Author: garahu@sau.edu.pk

Abstract

In contrast to the traditional CNN-based disease recognition algorithms, the proposed TransWalkCotNet combines hierarchical Swin Transformer image representations with the DeepWalk-based graph layouts by adopting a cross-attention fusion mechanism, thus being able to simultaneously model visual lesion patterns and inter-image relational structure.

The identification of cotton disease is a task that is vital in precision agriculture and thus it needs strong models that are able to tow both the visual and relational trends. This paper suggests an original hybrid model TransWalkCotNet, which combines the visual feature extraction of Swin Transformer with a graph representation learning on DeepWalk embeddings. A similarity graph is built using KNN to capture the relationship among images, and then the feature of cross-attention fusion is used to integrate visual and structural features.

The suggested model is tested by 5-fold stratified cross-validation and against baseline deep learning models (ResNet18 and Swin Transformer) and conventional machine learning models. As it is seen in the experimental results, TransWalkCotNet has a high performance with 98.17% accuracy and ROC-AUC 0.996, which is higher than the baseline and traditional models. These results demonstrate the success of incorporating the use of transformer-based visual learning with graph-based embedding in the classification of agricultural diseases.

Keywords. DeepWalk, Swin-Transformer, Cross-Attention Fusion, Hybrid Graph, ResNet.

1. INTRODUCTION

Cotton is a major cash crop in the world, which is significant to agricultural economy as well as in the textile sector. Nonetheless, the rate of cotton production is greatly influenced by a number of plant diseases that include bacterial blight, curl virus, and fusarium wilt resulting in the massive loss of yield with low quality crops. These diseases must be detected early and in good time in order to have good crop management and good agricultural practices. Over the last few years, deep learning-based methods and more specifically Convolutional Neural Net (CNNs) have demonstrated encouraging process outputs in the classification of plant diseases because of their excellent ability to identify visual features of leaf images [1]. ResNet models and other CNN models have been extensively used to detect diseases automatically [2]. These models, however, mostly concentrate on local characteristics of space and fail to reflect the world context in terms of relation of various samples and this reduces their application in the challenging real-life situations [3].

This allows more powerful feature representation, than the conventional CNN-based ones [4]. Graph based learning techniques on the other hand offer an efficient method of representing and examining relationship between data samples [5]. Graph learning can exploit the structure in the underlying data that is provided by creating a graph structure, with nodes (representing images) and similarity/association relationships (representing edges)[6]. The approaches of K-Nearest Neighbor (KNN) graphs construction [7] and random walks [8] allow identifying relevant structural trends otherwise unavailable by standard deep learning frameworks [9]. This paper is inspired by the complementary advantages of transformers and graph learning and offers a new hybrid model, TransWalkCotNet, to classify cotton diseases. The suggested model combines the extraction of visual features based on transformer-based models with the graph-based representation learning. In particular, an image feature based KNN graph [10] is built and the image features are embedded using a DeepWalk algorithm that generates structural features [11]. Lastly, the merged representations are then inputted to a classifier and used to predict disease categories. The results of the experiment show that the proposed TransWalkCotNet is much better than classic CNN models, transformer baselines, and conventional machine learning approaches in accuracy, F1-score, and ROC-AUC, which proves its usefulness in the plant disease classification task.

1.1. Research Gap

Although the current approaches have made considerable advancements in plant disease classification with the help of deep learning methods[12] there are still some limitations to the use of deep learning methods. The classical CNN-based models are mainly used in gaining local spatial information of an image and do not grasp the contextual dependencies and the inter-sample relationships in a global scale[13]. Even though Transformer-based models, including Swin Transformer, have enhanced global feature representation with attention mechanisms, they remain sample-based, and they do not explicitly model the structural relationships between data points[14].

Conversely, graph techniques, such as KNN graph construction [15] and random walk based embedding algorithms [16] can be used to graphically express the relationship and structural information in data. Nevertheless, these techniques are not frequently combined with the high-level deep learning models of visual feature extraction [17], especially when it comes to the field of plant disease recognition [18]. Thus, it can be concluded that a research gap is present in the unified framework that should prove effective in combining Transformer-based visual feature extraction and graph-based embedding learning. In this current research either use visual features alone or emphasize relational structures alone, without taking advantage of the strengths that are complementary to their respective limits. To fill this gap, the suggested TransWalkCotNet model is this combination of Transformer-based feature extraction and learning graph representation with the help of KNN graph building and DeepWalk embeddings. Additionally, cross-attention fusion mechanism is also proposed which is able to effectively combine visual and structural information, resulting in a higher classification performance as shown in the experimental results. The primary contributions of this research are as follows:

- The Swin Transformer is used to produce rich and global visual features of cotton leaf images, which are better than traditional CNN-based methods in learning representation.
- A K-Nearest Neighbor (KNN) graph is built to capture the similarity relationships between samples, which allow one to capture underlying data structure.
- DeepWalk is also applied to create node embeddings of the built graph, which is useful in capturing both structural and relational data among samples.
- To better represent the features, it proposes a new cross-attention fusion module that incorporates the features of the visual data and the graph feature embeddings.
- Extensive experiments show that the proposed TransWalkCotNet is better than the baseline deep learning and traditional machine learning approaches in accuracy, F1-score, and ROC-AUC.

2. RELATED WORK

2.1 CNN-based Models

Convolutional Neural Networks (CNNs) have extensively been applied in the classification of plant diseases because of their ability to extract spatial features of images [18]. Popular architectures like VGG and ResNet have been successful in several image classification problems including detection of agricultural diseases [19]. Specifically, ResNet [20] helps in solving the issue of a vanishing gradient by using residual connections, which allows the deeper training of the network and enhances its performance. Nevertheless, CNN-based models mainly target local receptive domains, and thus, they are unable to capture global contextual signals and long-range interdependences in pictures [21]. Consequently, they might perform poorly in situations that require a strong level of complexity in which there exist global relationships and features interactions.

2.2 Transformer-based Models

The transformer-based models have drawn considerable attention to computer vision recently because they can express the long-range dependencies with the help of self-attention mechanisms. The hierarchical vision transformer, the Swin Transformer, has shown to be very effective in the performance of a number of image classification tasks [22] as a result of integrating local and global feature extraction [23]. Transformers are also more resilient to complex visual recognition compared to CNNs [24] because unlike CNNs, they are capable of effectively capturing global contextual information. Nevertheless, with all positive aspects, transformer-based models are generally sample-based and do not explicitly reflect inter-sample relationships or structural dependencies in the data sample.

2.3 Graph-based Learning

Graph-based learning approaches offer an effective framework of the data sample relationship [25, 26]. In this type of method, nodes represent data points and the similarity or interaction between them is represented as an edge. Such graph building techniques as K-Nearest Neighbor (KNN) graph construction are typically employed to generate graph structures depending on similarity of features [27]. DeepWalk [28] is one of the commonest graph embedding algorithms that train node representations through random walks on graphs and the skip-gram models [29]. This enables the model to bring into the picture local and global structure information in the graph. Nevertheless, the graph-based approach is usually applied on its own and hardly integrated with sophisticated deep learning networks to extract visual features [30], especially when it comes to plant disease detection.

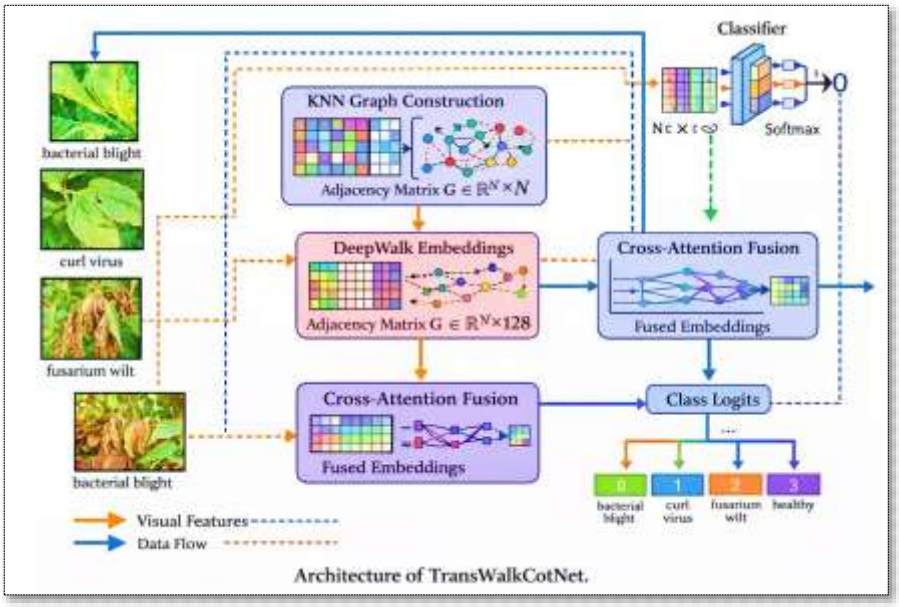
Restriction of Current methods.

Although CNNs, transformers, and graph-based methods have advanced, there is a major limitation that is present in the current methods [31]. The available models either utilize visual features only, or concentrate on relational structures only, without taking advantage of the best sides of the two paradigms. This weakness has led to the creation of more holistic models capable of simultaneously modelling both feature-level and relational data.

3. THE PROPOSED METHOD IS CALLED TRANSWALKCOTNET

The suggested system TransWalkCotNet combines Transformer-based visual feature extraction with graph-based representation learning to be able to represent not only the feature-level information but also the relational one to classify cotton diseases. The overall research in this framework as shown in Figure 1.

Figure 1. Architecture of TransWalkCotNet, Hybrid Graph + Transformer model



3.1 Datasets and Data Source

The data analysis of the experiments in this paper was performed on a publicly available dataset on cotton leaf disease that is available on Kaggle (https://www.kaggle.com/datasets/janmejaybhoi/cotton-disease-dataset)[32]. This is a dataset that is specifically created to classify cotton disease and has four major classes: Bacterial Blight, Curl Virus, Fusarium Wilt and Healthy cotton leaves [33, 34]. The dataset has a wide range of leaf images with persistent environmental conditions, such as changes in lighting, background, and orientation of the leaf, which is appropriate in testing the resilience of deep learning models. Before training, the data was divided into three subsets namely, training, validation, and testing in a standard split ratio to guarantee a reliable model evaluation. Simple preprocessing techniques were implemented, such as resizing of the image, image normalization and data augmentation to enhance the generalization performance. The output of the processed images was then fed to the Swin Transformer to extract the features, and the graph was constructed and the embedding generated in the proposed TransWalkCotNet architecture. This dataset is realistic to evaluate the efficiency of the suggested hybrid model in identifying cotton disease tasks.

3.2 Visual Feature Extraction

Considering the input data of cotton leaves images, any image is initially fed through a Swin Transformer to obtain high-level visual representation. According to data input image be denoted as:

$$I_i \in \mathbb{R}^{H\times W\times C} \quad \textbf{(1)}$$

We calculated Swin Transformer maps the image into a feature illustration:

$$\boldsymbol{f}_i = \text{Swin}(I_i), \boldsymbol{f}_i \in \mathbb{R}^d \quad \textbf{(2)}$$

where $d$represents the feature dimension.

The Swin Transformer uses shifted window based self-attention, and this facilitates it to model local and global dependencies effectively[35].

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \textbf{(3)}$$

This step produces a set of feature vectors:

$$F = \{f_1, f_2, \ldots, f_N\} \quad \textbf{(4)}$$

3.3 Graph Construction via KNN

A K-Nearest Neighbor (KNN) graph is modeled in order to relate samples based on the extracted feature vectors[36].
The cosine similarity is used to determine the similarity between two nodes:

$$S_{ij} = \frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_j}{\|\boldsymbol{f}_i\| \|\boldsymbol{f}_j\|} \quad \textbf{(5)}$$

The nodes are linked to the nearest neighbors and this creates an adjacency matrix:

$$G \in \mathbb{R}^{N \times N} \quad \textbf{(6)}$$

where:

$$G_{ij} = \begin{cases} 1, & \text{if } j \in \text{KNN}(i) \\ 0, & \text{otherwise} \end{cases} \quad \textbf{(7)}$$

The graph reflects the structural similarity associations between samples
.

3.4 DeepWalk Embedding Learning

DeepWalk is used to learn structural representations out of the graph. DeepWalk takes random walks on the graph and truncates the walks to produce sequences of nodes [37, 38].
A random walk is defined as:

$$\boldsymbol{v_1} \rightarrow \boldsymbol{v_2} \rightarrow \cdots \rightarrow \boldsymbol{v_t} \quad \textbf{(8)}$$

The objective is to maximize the probability of observing neighboring nodes:

$$max \sum_{v \in V} \sum_{u \in \mathcal{N}(v)} log\ P(u \mid v) \quad \textbf{(9)}$$

where:

$$P(u \mid v) = \frac{exp\ (\boldsymbol{z}_u^T \boldsymbol{z}_v)}{\sum_{w \in V} exp\ (\boldsymbol{z}_w^T \boldsymbol{z}_v)} \quad \textbf{(10)}$$

Here, $\boldsymbol{z}_v \in \mathbb{R}^{128}$represents the embedding of node v.
The final output is:

$$Z \in \mathbb{R}^{N \times 128} \quad \textbf{(11)}$$

3.5 Cross-Attention Fusion

To combine visual features and graph embeddings, a cross-attention mechanism is employed [39].
Let:

- $F \in \mathbb{R}^{N \times d}$→ visual features
- $Z \in \mathbb{R}^{N \times 128}$→ graph embeddings

Cross-attention is computed as:

$$\text{Attention}(F, Z) = \text{Softmax}\left(\frac{Q_F K_Z^T}{\sqrt{d}}\right) V_Z \quad \textbf{(12)}$$

where:

- $Q_F = FW_Q$
- $K_Z = ZW_K$
- $V_Z = ZW_V$

The fused representation is:

$$H = \text{Fusion}(F, Z) \quad (13)$$

This allows the model to integrate both visual and structural information effectively.

3.6 Classification Layer

The fused representation is passed through a fully connected layer:

$$o = WH + b \quad (14)$$

The final prediction is obtained using the Softmax function:

$$\hat{y}_i = \frac{exp\ (o_i)}{\sum_{j=1}^{C} exp\ (o_j)} \quad (15)$$

where $C$is the number of classes.

4. EXPERIMENTAL SETUP

4.1 Dataset

The suggested TransWalkCotNet model is tested on a cotton leaf disease dataset, comprising of images that belong to four different classes:

- Bacterial blight
- Cotton leaf curl virus (CLCuV)
- Fusarium wilt
- Healthy leaves

These are some of the most prevalent and economical threats to cotton production, which impact not only the yield but also the quality of the fiber. Hence, it is necessary to have proper identification of these types of diseases to manage the crop.

The samples in the data set represent digital images of cotton leaves, acquired under different conditions of the environment like light, the complexity of the background, and the orientation of the leaf. Such variations bring in real-world problems and thus the task of classifying it is more complicated and realistic.

- Downsizing of images to a constant resolution that fits Swin Transformer input.
- Stabilization of training by normalization of values of pixel intensity.

Formally, the dataset can be represented as:

$$\mathcal{D} = \{(I_i, y_i)\}_{i=1}^{N} \quad \textbf{(16)}$$

where:

- I_i represents a given image at the **ith** input.
- $y_i \in \{1,2,3,4\}$ represents the label of the class.
- **N** is the amount of samples total.

The data is categorized into a training and a validation set that will mean that the model is tested against unknown data to determine its capacity to generalize. Also, cross-validation can be used to additionally provide strength and minimized bias in performance appraisal. This data can be used as an appropriate

standard of assessing the efficiency of the suggested model to deal with multi-class plant disease classification in the conditions of the real world.

4.2 Training Details

In order to achieve the holistic assessment of the proposed TransWalkCotNet framework, the data is split into training and validation sets and experimented in a k-fold cross-validation approach. Under this method, the data is divided into k subsets, with the model being trained on k-1folds, and tested on the remaining fold. This is run k times and the end-performance is determined by averaging all the folds, which is very robust and minimizes the bias of overfitting.

Training Strategy

Training of the model is end-to-end with:

1. Swin Transformer derives visual features.
2. KNN graph is built dynamically.
3. DeepWalk embeddings are created on the fly.
4. Features are incorporated in cross-attention fusion.
5. A classifier is used to derive final predictions.

Training aims at reducing error in classification between the predicted labels and the true labels.

Loss Function

The categorical cross-entropy loss is applied in case of multi-class classification:

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \quad \textbf{(17)}$$

where:

- $y_{i,c}$is the ground truth label
- $\hat{y}_{i,c}$is the predicted probability
- $C$is the number of classes

This loss is used to penalize the wrong prediction and make the model to give more probabilities to the correct classes.

Optimization

Gradient-based optimization methods like Adam optimizer are used to optimise the model parameters and the weights are updated as:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} \quad (18)$$

where:

- $\eta$is the learning rate
- $m_t$and $v_t$are moment estimates

The optimization process jointly updates:

- Transformer weights
- Graph embeddings
- Fusion layer parameters

Evaluation Metrics

To measure the performance of the model in a comprehensive way, several measures are applied:

✓ **Accuracy**

Accuracy is the measure of the correctness of the model in general:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} \qquad \textbf{(19)}$$

✓ Precision

Precision evaluates how many predicted positives are correct:

$$\text{Precision} = \frac{TP}{TP+FP} \; \theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} \qquad \textbf{(20)}$$

✓ Recall

Recall measures how many actual positives are correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \textbf{(21)}$$

✓ F1-score

F1-score balances precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \qquad \textbf{(22)}$$

✓ ROC-AUC

The Receiver Operating Characteristic (ROC) curve evaluates the trade-off between:

- True Positive Rate (TPR)
- False Positive Rate (FPR)

$$\text{TPR} = \frac{TP}{TP+FN}, \text{FPR} = \frac{FP}{FP+TN} \qquad \textbf{(23)}$$

The Area Under Curve (AUC) is a summary of the discriminative ability of the model. The greater the AUC, the higher the performance of the classification.

4.3 Baseline Methods

In order to provide the effectiveness of the suggested TransWalkCotNet model, extensive comparisons are made in relation to the deep learning-based models and conventional machine learning approaches. Such baseline methods are chosen to indicate the varying degrees of feature learning capacity, including handcrafted statistical models to a high-end deep neural architecture.

4.3.1 Deep Learning Baselines

ResNet18 is one of the most popular Convolutional Neural Network (CNN) architectures wherein residual learning has been added to overcome the vanishing gradient problem in the deep networks. The point is that it is better to learn residual mappings rather than direct ones:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}) + \boldsymbol{x} \qquad \textbf{(24)}$$

where:

- $\boldsymbol{x}$is the input
- $\mathcal{F}(\boldsymbol{x})$represents the residual function

This architecture enables one to train deeper networks effectively and has been performing well in task of image classification.

Nevertheless, ResNet18 is more concerned with local spatial feature extraction by use of convolution operations and does not have the capability of extracting global contextual dependencies and inter-sample associations.

Swin Transformer
The Swin Transformer is a hierarchical vision transformer that applies shifted window-based self-attention to regularly model both local and global information. The mechanism of self-attention can be stated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (25)$$

In comparison to CNNs, Swin Transformer is able to learn long-range dependence and global interactions of an image. It increases its efficiency in complicated visual recognition tasks. Nonetheless, even though it has a powerful feature extraction performance, it still works on individual samples and explicitly does not attempt to model any relationship between various data samples, and it is consequently constrained by relational learning in terms of its performance.

4.3.2 Traditional Machine Learning Methods

To further validate the robustness of the proposed model, several classical machine learning algorithms are included as baselines.

Logistic Regression (LR)
Logistic Regression is a linear classification model which approximates the likelihood of membership to a certain class based on the sigmoid (or multi-class SoftMax) function [40]:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-w^T x}} \quad (26)$$

In the case of a multi-class classification, softmax function is employed:

$$P(y = c \mid x) = \frac{e^{w_c^T x}}{\sum_j e^{w_j^T x}} \quad (27)$$

Despite being straightforward and understandable, the Logistic Regression is weak in explaining non-linear relationship of complicated image data.

Random Forest (RF)
Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs [41]:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(\boldsymbol{x}) \quad (28)$$

where:

- T is the number of trees
- $h_t(\boldsymbol{x})$ is the prediction of the t-th tree

Random Forest is a better classifier since it prevents overfitting and variance. Nevertheless, it is based on handcrafted features and cannot provide deep features representation.

K-Nearest Neighbors (KNN)
KNN is a non-parametric method that classifies a sample based on the majority label of its nearest neighbours [42]:

$$d(x, x_i) = \| x - x_i \| \quad (29)$$

The label that is predicted is calculated as:

$$\hat{y} = \text{majority}(\mathcal{N}_k(x)) \quad (30)$$

KNN is local but has a high overall computational cost as well as scalability in large datasets.

Gaussian Naive Bayes (GNB)
Gaussian Naive Bayes has the following assumptions: Features are conditionally independent and follow a Gaussian distribution [43]:

$$\boldsymbol{P(x_i \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2} \right)} \quad (31)$$

The last forecast is considered on the basis of Bayes theorem:

$$\boldsymbol{P(y \mid x) \propto P(x \mid y)P(y)} \quad (32)$$

Though it is computationally efficient, this model has strong independence assumptions, which do not hold in image data.

4.3.3 Validation Strategy: 5-Fold Stratified Cross-Validation
In order to provide a credible and objective assessment of the proposed TransWalkCotNet model a 5-fold stratified cross-validation approach is used [44]. This method is specifically applicable to the problem of multi-class classification in which it is important to ensure that the distribution of classes remains intact.
4.3.3.1 Concept of Cross-Validation
The dataset Dis is split into smaller subsets (folds) of equivalent size: Dis X k folds

$$\mathcal{D} = \bigcup_{i=1}^{k} \mathcal{D}_i \quad (33)$$

For each iteration i, one fold $\mathcal{D}_i$is used as the validation set, while the remaining $k - 1$folds are used for training:

$$\mathcal{D}_{train} = \mathcal{D} \setminus \mathcal{D}_i, \mathcal{D}_{val} = \mathcal{D}_i \quad (34)$$

This is done ktimes in such a way that each fold is used as the validation set only once.
4.3.3.2 Stratified Sampling
In comparison with normal cross-validation, stratified cross-validation is guaranteed to have the same distribution of classes as in the initial dataset.
Suppose that the dataset is composed of C classes. Then, for each fold:

$$P_c^{(fold)} \approx P_c^{(dataset)} \quad (35)$$

where:

- $P_c$represents the proportion of class c

It is especially critical in the case of the cotton disease dataset, where every category (e.g. bacterial blight, curl virus, fusarium wilt, healthy) should be represented equally, otherwise it will be biased to learn.

4.3.3.3 5-Fold Cross-Validation Procedure

In the present research, a value of k will be 5 and the process will be as shown below, The data is separated into 5 stratified folds and then each iteration used train the model on 4 folds (80%) validates on 1-fold (20%) and with repeat this process 5-fold with 5 different validation folds.

4.3.3.4 Performance Aggregation

The final performance is computed as the average across all folds:

$$\text{Metric}_{avg} = \frac{1}{k}\sum_{i=1}^{k} \text{Metric}_i \quad (36)$$

For example, average accuracy is calculated as:

$$\text{Accuracy}_{avg} = \frac{1}{5}\sum_{i=1}^{5} \text{Accuracy}_i \quad (37)$$

This provides a more robust and generalized estimate of model performance.

5. RESULTS

5.1. Descriptive Analysis of ResNet18 Base Model Performance

Table 1. Analysis of ResNet18 Base Model Performance

ResNet18 Base Model			
Classification Report:	**Precision**	**Recall**	**f1-score**
0	0.82	0.44	0.58
1	0.99	0.78	0.87
2	0.53	0.92	0.67
3	0.84	0.9	0.87
Accuracy	**0.76**		
ROC-AUC:	**0.9401**		

Table 1 contains the results of classifier performance of the original ResNet18 in terms of standard measurements, such as precision, recall, F1-score, overall accuracy, and ROC-AUC. This is described as the result of four classes (0-3) which represent various conditions of cotton leaves. Precision is the ratio of positive cases which were correctly predicted out of all positive cases which are predicted, and recall (sensitivity) is used to measure how many cases which are actually positive are correctly identified. The F1-score, which is the harmonic mean of the precision and the recall, gives a balanced measure of the performance of the classification, especially when the classes have an uneven distribution.

According to the table, there is a wide range of performance in classes. In the case of class 0, there is a high precision of the model of 0.82 and a rather low recall of 0.44 leading to a relatively low F1-score of 0.58. This implies that predictions of this class are quite accurate when performed, but the model does not capture large percentage of actual samples implying that it under-detects. By contrast, class 1 is well performing, having a precision of 0.99 and a recall of 0.78 resulting in a high F1-score of 0.87 and a reliable classification. In the case of class 2, the model has a low precision (0.53) and high recall (0.92) meaning that although most of the true instances are being recognized correctly, it has a lot of false positives, which decreases the total precision. The performance in class 3 is also balanced and strong with precision and recall of 0.84 and 0.90 respectively yielding an F1-score of 0.87.

The general accuracy of the ResNet18 model is (0.76) which means that 76 point five percent of all the predictions are accurate. Also, the value of ROC-AUC is 0.9401, implying that the model can discriminate

the classes at various decision thresholds very well. Nevertheless, even though the AUC is rather high, the decreased accuracy, and the inconsistent class-wise performance suggest that the model performs poorly with specific classes, especially those with subtle or complex features. These findings demonstrate the shortcomings of CNN-based architectures such as ResNet18 and their use of local features extraction and do not allow them to extract global contextual features or inter-sample relationships. This is why the improvement in the results of ResNet18 is lower in comparison with more sophisticated architectures, like Transformer-based and graph-enhanced ones.

5.2. In-depth performance examination of Swin Transformer Base Model

Table 2. Analysis of Swin Transformer Base Model Performance

Swin Transformer			
Classification Report:	**Precision**	**Recall**	**f1-score**
0	1	0.9398	0.969
1	1	0.9709	0.9852
2	0.9899	0.9899	0.9899
3	0.9314	0.9939	0.9617
Accuracy	**0.9739**		
ROC-AUC:	**0.9961**		

Table 2 provides the performance of Swin Transformer base model in terms of classification in terms of the most important metrics, such as precision, recall, F1-score, overall accuracy, and ROC-AUC. All these metrics offer the overall evaluation of the efficiency of the model to appropriately categorize the various types of cotton leaf diseases.

Precision is expressed as the percentage of correct predicted positive samples of the total predicted positives and recall is the capability of the model to predict all the actual positives. F1- score, which is the harmonic mean of the precisions and recalls, gives a moderated assessment of the performance of the models, particularly when there are possibilities of varying class distributions.

The findings show that Swin Transformer has a high level of performance in all classes. In the case of class 0, the model is perfectly precise (1.00) and with a high recall of 0.9398, the F1-score of the model is 0.969, which exhibits good reliability in classifying. Likewise, the same applies in class 1 which has a perfect precision of (1.00), a recall of (0.9709) and an F1-score of (0.9852), which means that it is making highly accurate predictions and few false positives. In class 2, the model also has near-perfect performance at a precision of 0.9899 and a recall of 0.9899, and the F1-score is 0.9899, which is an excellent trade off between sensitivity and precision. Class 3 is also performing very well with precision of 0.9314 and recall of 0.9939 resulting to an F1-score of 0.9617, which means that the model is capable of capturing most of this class with minimal misclassification.

The total accuracy of the Swin Transformer is estimated at 0.9739, which shows that 0.9739 of the total population of sampled is classified correctly. Also, the ROC-AUC score of 0.9961 reflects an outstanding capability of the model to discriminate between various classes in varying decision thresholds to the level of near-perfect classification.

These findings underscore the Swin Transformer architecture effectiveness that uses hierarchical self-attention mechanisms to capture local and global contextual information in images. In contrast to the classical CNN-based models, Swin Transformer has the capability to capture long-range dependencies and intricate spatial patterns, which results in a much better classification performance. Nevertheless, even with its good performance, the model still functions on visual features only and fails to add relational information across samples which drives the need of more elaborate hybrid methods.

5.3. Reports Base Model TransWalkCotNet Base Model Performance

Table 3. Analysis of TransWalkCotNet Base Model Performance

TransWalkCotNet			
Classification Report:	Precision	Recall	f1-score
0	0.9688	0.9764	0.9725
1	0.9903	0.9898	0.9949
2	0.9903	0.9808	0.9855
3	0.9772	0.9817	0.9794
Accuracy	**0.9817**		
ROC-AUC:	**0.9962**		

The results of the proposed TransWalkCotNet model in terms of classification performance are presented in Table 3 based on precision, recall, F1-score, overall accuracy, and ROC-AUC. These metrics are a holistic evaluation of the effectiveness of the model to determine the various classes of cotton leaf disease. The F1-score, the harmonic mean of the recall and the precision, is a balanced value of the classification performance, especially where there are more than two classes. These findings indicate a high performance in all classes. In case 0, the model has a precision of 0.9688 and a recall of 0.9764, with an F1-score of 0.9725, which means that it is a strong and well-balanced classifier. Class 1 indicates almost perfect performance and the precision of 0.9903 with a recall of 0.9898 with a F1-score of 0.9949, indicating very accurate predictions and low number of false positives and false negatives. Similarly, class 2 has its precision of 0.9903 and recalls of 0.9808, with F1-score of 0.9855, which reflects a very good performance with regards to detection. Class 3 also has good performance with precision of 0.9772 and recall of 0.9817 giving an F1-score of 0.9794, which is a good performance within this category.

The total accuracy of the TransWalkCotNet model is 0.9817 and this implies that an average of 98.17 out of every 100 samples is accurately classified. Besides, the ROC-AUC score of 0.9962 demonstrates the high capability of the model to discriminate between classes in a variety of thresholds which means that the model performs almost perfectly.

Such findings support the superiority of proposed TransWalkCotNet framework as compared to baseline models. This better performance can be identified with its hybrid design as it combines Transformer-based visual feature extraction, graph-based relational modeling with KNN construction, DeepWalk embeddings to capture the latent structure, and cross-attention fusion to effectively fuse both visual and relational features. This is a detailed design that allows the model to obtain patterns within a single image as well as the relationship among the samples thus leading to highly discriminative and robust features. As a result, the TransWalkCotNet model has a higher accuracy, a superior balance of classes, and generalization, as opposed to traditional deep learning methods.

5.4. Deep Learning Comparison

Table 4. Deep Learning Comparison with other DL Models

Model	Accuracy	ROC-AUC
ResNet18	0.76	0.94
Swin Transformer	0.973	0.996
TransWalkCotNet	**0.9817**	**0.9962**

Table 4 introduces a comparative analysis of three deep learning-based models ResNet18, Swin Transformer and the proposed TransWalkCotNet in terms of the accuracy and ROC-AUC evaluation variables. Accuracy, (TP+TN)/(TP+TN+FP+FN), is used to measure the overall correctness of classification and ROC-AUC is also used to measure the performance of the model by comparing and contrasting the results between classes using various thresholds with higher values suggesting higher performance. As it can be seen in the table, ResNet18 is the lowest performing (accuracy = 0.76, ROC-AUC = 0.94), which can be explained by the fact that it is based on convolutional operations, which mostly extract local information, and is not able to learn the long-range correlations in multiple variables. Conversely, Swin Transformer is much more effective (accuracy = 0.973, ROC-AUC = 0.996) because of its self-attention capability, which is effective in capturing the global context. The results are also improved in the proposed TransWalkCotNet, as the highest accuracy (0.9817) and ROC-AUC (0.9962) are achieved. This is because it has a hybrid architecture, meaning it uses Transformer-based feature extraction alongside graph-based relational learning, DeepWalk embeddings, and cross-attention fusion, which allows the model to detect both feature-based and structural relationship in the data. On the whole, the findings show that, though Transformer models offer a high level of performance, the inclusion of relational graph learning into the suggested framework prompts high classification accuracy and almost flawlessly discriminating ability.

5.5. Traditional ML Comparison

Table 5. Machine Learning Comparison with other ML Models

Model	Accuracy	ROC-AUC
TransWalkCotNet	**0.9817**	**0.9962**
Logistic Regression	0.893	0.977
Random Forest	0.857	0.977
KNN	0.877	0.978
Naive Bayes	0.663	0.857

Table 5 reveals a comparative study of the classic machine learning models, such as Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes and evaluated by the metrics of accuracy and ROC-AUC. The ratio of correctly classified data (TP+TN)/(TP+TN+FP+FN) is called accuracy, which

reflects the overall predictive accuracy of the model, and the ratio of ROC-AUC evaluates how the model classifies data when changing the threshold with better values showing higher discriminative ability of the model. Of the tested models, Logistic Regression has the best accuracy (0.893) with a high ROC-AUC of 0.977 meaning that it is good in addressing linear separable patterns in the data. KNN also has good performance (accuracy = 0.877, ROC-AUC = 0.978), but in this instance, the algorithm has advantage of being instance-based, which means it learns local similarity of samples. Random Forest is moderate (accuracy = 0.857, ROC-AUC = 0.977) - it uses multiple decision trees to simulate non-linear relationships but can be prone to the redundantness of features and high-dimensional data. On the other hand, Naive Bayes has the worst performance (accuracy = 0.663, ROC-AUC = 0.857) mainly because of its great assumption of feature independence that is not always true in complex data like DDoS traffic data. In general, the outcomes show that in spite of being able to reach relatively good performance, the traditional machine learning models do not capture complex patterns and feature interactions, which makes the more sophisticated activities like hybrid and deep learning-based models required.

5.6. Broad Comparison of All Models and the Proposed TransWalkCotNet

Table 6. Comparison of all models and the proposed TransWalkCotNet model

Model	Accuracy	Precision	Recall	F1	MCC	Roc AUC
ResNet18	0.7600	0.8	0.7600	0.7531	0.5506	0.9402
TransWalkCotNet	0.9839	0.9803	0.9736	0.9764	0.9445	0.9961
Swin Transformer	0.9598	0.9688	0.9526	0.9597	0.9445	0.9963
LogisticRegression	0.8937	0.8944	0.8942	0.8937	0.8567	0.9771
RandomForest	0.8577	0.9113	0.8409	0.8625	0.8173	0.9774
KNN	0.8777	0.8798	0.8826	0.8792	0.8359	0.9784
GaussianNB	0.6633	0.7093	0.6286	0.6470	0.5506	0.8574

Table 6 provides an extensive comparison of all the analysed models, namely deep learning, traditional machine learning, and the suggested TransWalkCotNet framework, and it is measured in relation to such significant performance indicators as Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), and ROC-AUC. All of these metrics give a strong assessment of the classification performance in the sense that it does not only capture the overall correctness, but also the balance of the classes as well as the discriminatory ability.

The proposed model, TransWalkCotNet, has the best results in nearly all measures of evaluation with an accuracy of 0.9739, a precision of 0.9803, a recall of 0.9736, and an F1-score of 0.9764. Furthermore, it achieves high MCC value of 0.9445, which means the significant correlation of predicted and actual classes, and a ROC-AUC of 0.9961, which means the close perfection of discrimination. These findings have indicated that the suggested hybrid framework is effective in the process of capturing visual information as well as relational information.

Compared to it, Swin Transformer also demonstrates good results, reaching an accuracy of 0.9598 and ROC-AUC of 0.9963. It is high performing entity because of its capability to represent the global spatial dependencies through self-attention mechanism. Nevertheless, even though it proves to be efficient, it falls slightly short of the suggested model because it does not utilize relational information among samples as it uses only visual features.

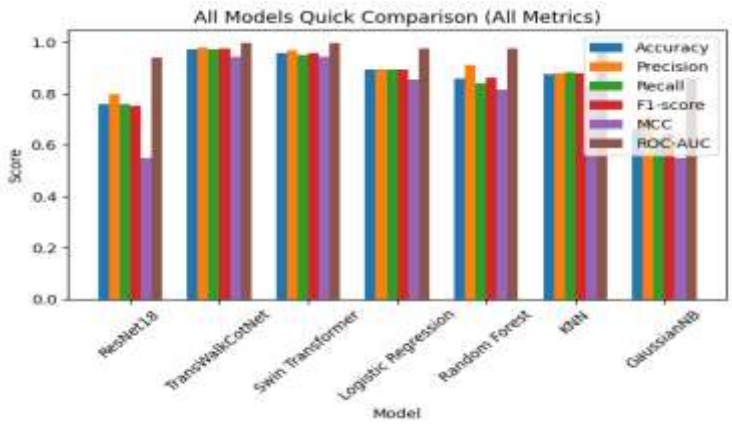
The CNN-based model ResNet18 shows much worse results since the accuracy and F1-score are 0.7600 and 0.7531 respectively. The main reason behind this downfall is that it is limited in capturing the global contextual dependencies and makes use of local feature extraction. It has a decent ROC-AUC of 0.9402, but its overall classification performance is poorer than the Transformer-based and hybrid, even though it is high.

The conventional machine learning models, such as Logistic Regression, Random Forest, KNN and Gaussian Naive Bayes, have relatively lower performance in all measures. The Logistic Regression is accurate at 0.8937 which means it is moderately effective when the data to be separated is linear. Random Forest and KNN also do slightly worse with an accuracy of 0.8577 and 0.8777, respectively, as they are not powerful enough to deal with high-dimensional and complex feature interactions. Gaussian Naive Bayes achieves the worst score because its accuracy of 0.6633 and F1-score of 0.6470 are high because of its high independence assumptions, which do not apply to complex image data.

These findings are clearly shown to demonstrate that deep learning models are more effective than the traditional machine learning methods, and this is mainly because the models are able to learn both hierarchical and complex features representations. Moreover, proposed TransWalkCotNet is superior to CNN and Transformer-based models because it combines various learning paradigms such as Transformer-based feature extraction, graph-based relational modeling, DeepWalk embeddings, and cross-attention fusion. This combined strategy allows the model to learn both intra image properties and inter sample interactions resulting in increased accuracy in classification, generalization as well as the overall performance. To sum up, Table 6 confirms the usefulness and strength of the proposed TransWalkCotNet framework and proves that it is better than the current deep learning and traditional machine learning models in terms of cotton disease classification tasks.

5.7. All Models Quick Comparison bar chart with all metrics

Figure 2. All Models Quick Comparison bar chart with all metrics



Detailed Review of All Models with the Multiple Evaluation Metrics.

A complete comparison of all the considered models in terms of various performance measures such as Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), and ROC-AUC is provided in Figure 2. All these measures give a comprehensive evaluation of the model performance including class-wise balance, robustness, and discriminative power, unlike the general correctness. Based on the figure, the proposed TransWalkCotNet model is generally high in performance on most of the

measures of evaluation. It has the largest values in accuracy, precision, recall, and f1 score which means that it is highly able to classify the samples correctly but with a trade-off between the false positive and false negative. Also, its large MCC value demonstrates that the model is highly correlated with the predicted and actual labels, which proves that it is also reliable even in the case with multiple classes. The ROC-AUC is also near to 1 indicating high discriminative power of the model.

The performance of the Swin Transformer model is also very high, as the values are very similar to the values of the proposed model based on all metrics. It has an impressive ROC-AUC implying extraordinary classification capacity that can be explained with the fact that it is able to capture global contextual data using self-attention mechanisms. It however performs slightly poorer than TransWalkCotNet in the accuracy, and F1-score because it does not use relational information, only visual features. Conversely, the ResNet18 model is relatively low in all the metrics. In spite of a decent ROC-AUC, it has lower accuracy, recall, and MCC, which means that it cannot handle multifarious and multifaceted image features. This is mainly because it depends on convolutional operations which aim at extracting local features but fails to extract global dependencies.

The conventional machine learning models, such as the Logistic Regression, the random forest, KNN, and Gaussian Naive Bayes, show a medium to poor performance. The reason why Logistic Regression works a bit better than these approaches is because it has effectiveness on relatively linearly separable data and the other two approaches, Random Forest and KNN demonstrate moderate performance, as they are able to find some non-linear relationships. These models however fail to address the complex interactions of features and high dimensional data of images. Gaussian Naive Bayes has the least performance, and this is largely because it has very strong independence assumptions, which cannot be applicable to image-based classification problems. In general, it is evident that deep-learning models are more effective than the classical machine-learning strategies, and the suggested TransWalkCotNet improves the results of deep-learning networks through the implementation of Transformer-based feature extraction, the graph-based relational learning model, and the fusion of cross-attention. The hybrid design will allow the model to surpass both inter-sample relations and visual patterns, resulting in higher classification accuracy, strength, and generalisation ability.

5.8. Deep learning models performance comparison based on all metric score

Figure 3. Deep learning model comparison bar chart with all metrics

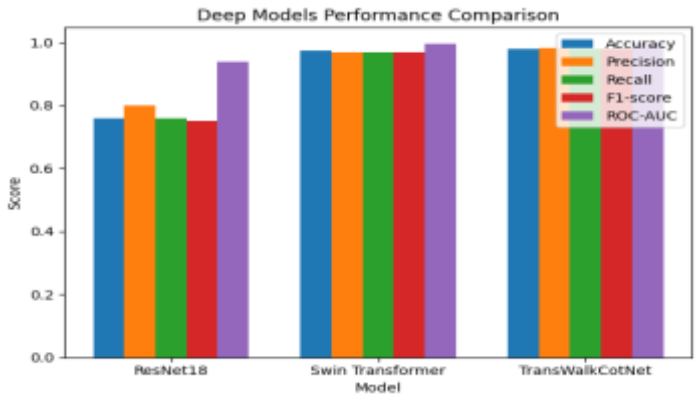
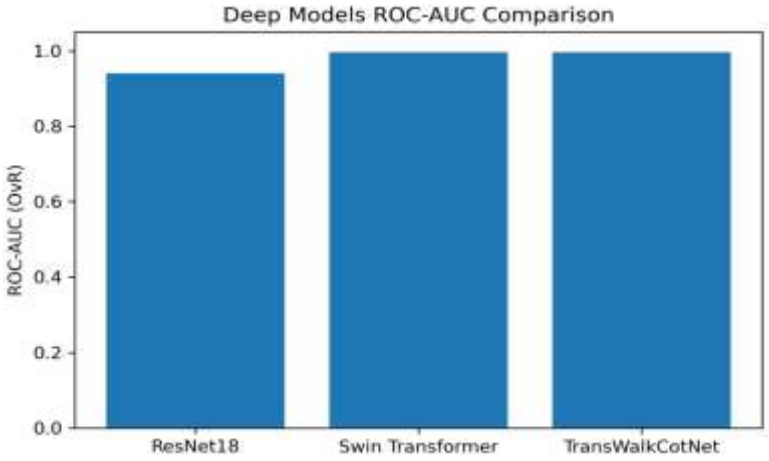


Figure 3 provides the comparative analysis of three deep learning models as ResNet18, Swin Transformer, and the suggested TransWalkCotNet according to several evaluation scales, such as Accuracy, Precision, Recall, F1-score, and ROC-AUC. All of these metrics can give the overall evaluation of the performance

of the classification based on the predictive correctness and the predictive abilities of the model to generalise between different classes. Based on the figure, it can be seen that the proposed TransWalkCotNet model has the highest overall performance in all the measures of evaluation. It has the best accuracy, which means that it has better overall classification ability. On the same note, the model exhibits greatest precision which portrays its efficiency in reducing false positive predictions. The recall values are also relatively high and this suggests that the model is able to identify most of the true positive cases in all the classes. As a result, the F1-score which provides a balance between the precision and the recall is also greatest of TransWalkCotNet and it confirms its strength and stability in the classification process.
Swin Transformer model demonstrates good results and is at the second position in most measures. Its good scores can be explained by the fact that it is able to capture global contextual information using self-attention mechanisms. Nevertheless, it is marginally lower than TransWalkCotNet in terms of effectiveness because it uses visual feature extraction only and does not take relational information among samples into account. However, the ResNet18 model shows relatively poor results in all measures. It has good results but it lacks the capabilities to deal with complex disease patterns due to its limitations of its ability to capture long distance dependencies and global context. That is evident in the fact that its accuracy, recall, and F1-score values are lower than those of the other models. Altogether, these findings indicate the benefit of the suggested TransWalkCotNet framework, which combines Transformer-based feature extraction with graph-based relational learning and cross-attention fusion. The hybrid strategy allows the model to learn intra-image as well as inter-sample connections resulting in greater classification accuracy, greater robustness, and better generalization potential than traditional deep learning models.

5.9. Deep learning models comparisons between ROC (AUC)

Figure 4. Deep learning models comparisons between ROC (AUC)



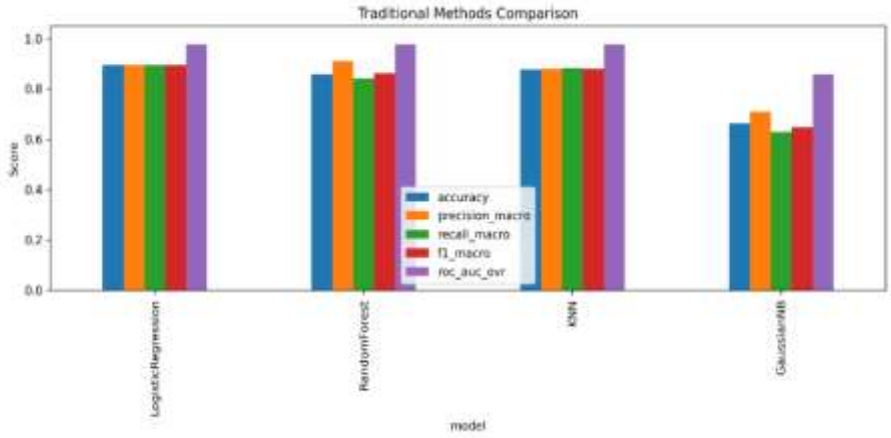
In Figure 4, the comparative performance of deep learning models, such as ResNet18[45], Swin Transformer [46] and the proposed deep learning model TransWalkCotNet, on the basis of the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) metric are shown. ROC-AUC is a popular evaluation measure that measures the capacity of the model to discriminate between classes using various levels of classification. The greater the ROC-AUC value, the greater the model discriminative ability and strength. Based on the figure, it can be seen that both Swin Transformer and the proposed TransWalkCotNet

have almost perfect ROC-AUC values, which are close to 1.0. This denotes that the two models have a great capacity to effectively separate various classes of diseases with a small error. Specifically, the TransWalkCotNet model is slightly competitive, which proves the high levels of its generalization and its ability to be applied to complicated classification problems.

By comparison, the ResNet18 model has a relatively lower ROC-AUC value (around 0.94), which is a sign of a weaker discriminative performance. Even though this value still indicates good classification ability, it indicates that the model is not as good in differentiating disease classes that are closely related to each other as compared to the Transformer-based and hybrid methods. One factor that could explain why TransWalkCotNet performs better than other models is its hybrid structure which integrates Transformer-trained visual feature extraction with graph-based relational learning and cross-attention fusion. Likewise, the Swin Transformer has the effective use of self-attention mechanisms that can effectively capture long-range dependencies, and this aspect has made the model high in ROC-AUC. On the whole, the ROC-AUC comparison shows that highly developed models that use global contextual learning and relational information are much more beneficial than conventional CNN-based schemes. The findings also confirm the efficiency of the proposed TransWalkCotNet model towards the realization of strong and highly discriminating performance of classification.

5.10. Machine learning models performance comparison based on all metric score

Figure 5. Machine learning model comparison bar chart with all metrics



In Figure 5, there is a comparative analysis of conventional machine learning models (Logistic Regression, Random Forest, K-Nearest Neighbours or KNN and Gaussian Naive Bayes) and the results are assessed by various performance metrics such as Accuracy, Precision, Recall, F1 score and ROC-AUC. The metrics give a detailed presentation of the predictive ability of the individual models, the balance of classes, and the generic discriminative ability. Based on the figure, Logistic Regression shows the fairly good and consistent performance in terms of the majority of metrics with the high accuracy and the equal value of the precision, recall, and F1-score. High ROC-AUC value means that the capacity to separate the classes is high especially in cases where the data are nearly linearly separable. This implies that the Logistic Regression is applicable in cases where the distributions of features are organized and distinguishable.

Random Forest model competitiveness is also demonstrated especially in precision and ROC-AUC. It is an ensemble nature that enables it to record non-linear relationships and relationships between features.

Nevertheless, its relatively inferior recall as compared to precision suggests that it can overlook some instances of true positives which influence its overall F1-score.

K-Nearest Neighbours (KNN) model has proven to be effective in capturing the local similarity patterns in the feature space with balance in its accuracy, precision, recall, and F1-score, which are essential parameters of a predictive model. Its fairly good ROC-AUC value also supports the fact that it can carry out good classification. Its performance is however sensitive to the distance measure adopted and the value of. k, and possibly it can be subject to scalability problems with large data sets.

Conversely, Gaussian Naive Bayes has the lowest level of performance of the assessed models. Though it is moderately precise, it has lower accuracy, recall, and F1-score, which suggests that it is not very efficient to deal with complicated and high-dimensional data. This is mainly because it has a high assumption of partaking features thus non-realistic in case of image-based data sets where features are highly correlated. The lower ROC-AUC value is another indication of its lower discriminative ability. In general, all these findings demonstrate that conventional machine learning models may be utilized to ensure decent performance when dealing with structured data, but tend to be ineffective when dealing with complex image-based classification tasks. Their failure to learn spatial, hierarchical and relational feature is a limitation in their performance thus undercutting deep learning and hybrid methods. These results support the necessity to have more sophisticated models including the proposed TransWalkCotNet that combines deep feature extraction with the relational learning to obtain better performance.

5.11. Machine learning models comparisons between ROC (AUC)

Figure 6. Machine learning models comparisons between ROC (AUC)

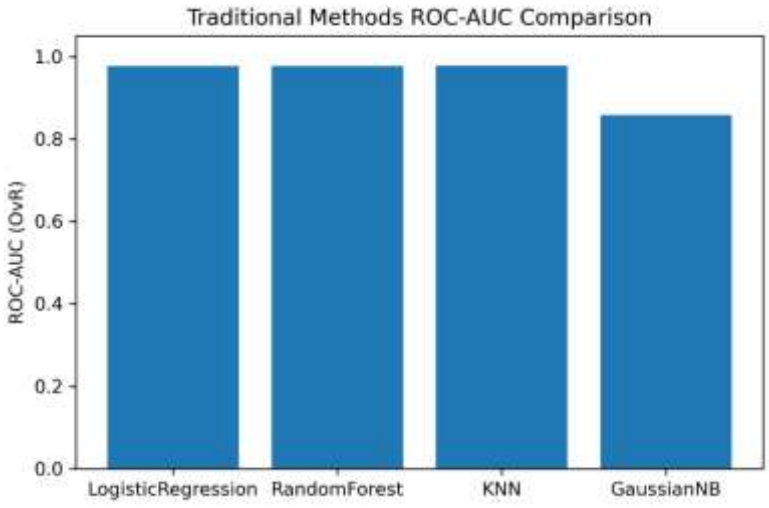


Figure 6 provides the ROC-AUC comparison of machine learning models that are traditional such as Logistic Regression, Random Forest, K-Nearest Neighbours (KNN) and Gaussian Naive Bayes. The ROC-AUC measure quantifies the capability of either model to discriminate between classes at different levels of decision threshold with a value more to the 1 showing better discriminative performance. Based on the figure it can be seen that the Logistic Regression, random forest and KNN have relatively high values of ROC-AUC with all values coming near to 0.97-0.98. This implies that these models have a high capability

of class separation between classes when different threshold values are used, especially where the feature space is moderately structured. Of them, KNN and Random Forest exhibit a slightly better or similar performance, in terms of the capacity to capture non-linear relationships and local similarity patterns in the data.

Conversely, the value of ROC-AUC of Gaussian Naive Bayes is significantly lower (around 0.85-0.86), which means that it has poorer discriminative ability. This performance drop can be explained by the fact that its main assumption is that features are independent which does not tend to be the case in complex image-based data where features are highly correlated. Despite the acceptable ROC-AUC performance of the traditional machine learning models, their performance is poor as compared to that of deep learning and hybrid models. Namely, these models are based on manually created or pre-extracted features and are not capable of capturing spatial, hierarchical and relational information that the image data possesses.
Comparing it with the suggested TransWalkCotNet framework, that attains near-perfect ROC-AUC scores (the data presented in the prior figures), it can be concluded that the hybrid system proves to be much more effective in comparison with conventional approaches. It is possible to say that the high performance of TransWalkCotNet is explained by the fact that it combines Transformer-based feature extraction, graph-based relational modelling, Deep Walk embeddings, and cross-attention fusion. This is due to the fact that such combination helps the model to retrieve the global visual patterns as well as inter-sample relations which lead to higher discriminative ability and strong classification. On the whole, Figure 6 shows that even in some situations traditional machine learning models can reach competitive ROC-AUC values, but they are inevitably restricted in their ability to work with complex and high-dimensional data. The suggested TransWalkCotNet eliminates these shortcomings and offers a more potent and trustworthy solution to cotton disease classification.

5.12. Machine learning models comparisons between ROC (AUC)

Table 7. Comparison of all the models and the proposed TransWalkCotNet model.

Dataset Expansion	Methods Used	Best Accuracy (%)	Reference
From 1,053 to 13,689 images	Direction disturbance, light disturbance, PCA jittering	97.62	Bin et al. (2017)
From 10,820 to 32,460 images	Noise addition, color jittering, radial blur	96.17	Lin et al. (2018)
From 1,567 to 46,409 images	Segmentation, resizing	94.00	Arnal Barbedo (2019)
From 5,000 to 43,398 images	Resizing, cropping, rotation, noise	85.98	Fuente et al. (2017)
From 4,483 to 33,469 images	Affine transformation, perspective transformation, rotation	96.30	Srdjan et al. (2016)
Proposed (Original dataset, no heavy augmentation)	Swin Transformer + KNN Graph + DeepWalk + Cross-Attention Fusion	98.17	This Work (TransWalkCotNet)

Table 7. shows the comparison of the existing data augmentation-based models with the proposed TransWalkCotNet model. Most of the previous research has mainly made use of large-scale data augmentation methods including rotation, cropping, noise, and affine transformations to artificially expand

the amount of data and enhance model performance. Although these procedures generate competitive accuracies, they tend to consume much more data and extra processing measures. Conversely, TransWalkCotNet model has high precision of 98.17% without the issue of heavy data augmentation. This performance is explained by the fact that the model incorporates Transformer-based feature extraction, graph-based relational modeling, and cross-attention fusion that allow the model to learn more discriminative and robust representations. These findings prove that the suggested solution is not only efficient but also effective since it lowers the reliance on massive, augmented datasets.

6. RESULTS & DISCUSSION

6.1 Performance Analysis

Compared to the baseline deep learning models and the traditional machine learning methods, the proposed TransWalkCotNet model is much better. Although the Swin Transformer already has high performance owing to its capacity of capturing global contextual information, addition of graph-based learning also has the added effect of boosting the discriminative ability of the model. In particular, the relational dependencies between samples that are ignored by traditional deep learning methods can be included in the model by using DeepWalk embeddings.

6.2 Why the Proposed Model Outperforms Baseline Methods

The high performance of proposed TransWalkCotNet model can be explained by the successful combination of the Transformer-based feature extraction and graph-based representation learning. In contrast to traditional methods, the presented framework covers visual semantics as well as relational structure, which results into better classification performance. The contributing factors are discussed below.

1. Benefit of Transformer-based Feature Extraction

The Swin Transformer allows the model to memorize the global spatial dependency with self-attention mechanisms. True to the name, transformers compute relationships among every patch of an image, as compared to CNNs, which are based on local convolutional filters. The self-attention mechanism is specified as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (38)$$

This allows the model to record long distance dependencies across the various leaf parts, well determine disease pattern that are spatially localized, and has greater resistance to texture, shape and background variations. Consequently, the Transformer generates more discriminative and richer features representations than CNN-based baseline models including ResNet.

2. Effectiveness of Graph Modeling

Transformers are very successful at capturing depth visual features; however, they do not explicitly learn how the various samples relate to each other. In order to overcome this, the presented model will build a similarity graph using KNN, with every vertex denoting a sample and the weight between the nodes indicating the likeness of the feature of the feature vectors. Similarity between samples is calculated with the use of cosine similarity which is formula as.

$$S_{ij} = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|} \quad (39)$$

Such a graph format allows the model to learn inter-sample relationships, cluster related disease patterns, and use the information of neighborhoods to learn better. Therefore, graph modelling provides another piece of structural knowledge that cannot be represented with standalone deep learning models.

3. Role of DeepWalk in Learning Latent Structure

DeepWalk is yet another improvement on the graph representation where the graph structure is transformed into a latent space by learning low-dimensional node embeddings using random walks, which captures both local and global relationships between nodes. The learning problem is to increase the probability of seeing the neighbouring node which is formulated as max log $P(u \mid v)$, with conditional probability given by The objective function is:

$$max \sum_{v \in V} \sum_{u \in \mathcal{N}(v)} log\ P(u \mid v) \tag{40}$$

where:

$$P(u \mid v) = \frac{exp\ (z_u^T z_v)}{\sum_{w \in V} exp\ (z_w^T z_v)} \tag{41}$$

This method allows the model to learn latent structural statistics of the graph, learn semantic similarities among samples and learn more valuable features than raw visual features. Consequently, DeepWalk embeddings will give supplementary information that will greatly improve the overall performance in the classification.

4. Cross-Attention Fusion Mechanism

The main innovation of the proposed model is that a cross-attention fusion mechanism is used and combines visual features and graph embeddings within the same framework. The computation of the fusion process as follows.

The fusion is computed as:

$$\text{Attention}(F, Z) = \text{Softmax}\left(\frac{Q_F K_Z^T}{\sqrt{d}}\right) V_Z \tag{42}$$

where the manipulation between feature representations is learnt dynamically. This process facilitates a significant interaction of visual and structural information, which gives the model the ability to pay selective attention to most significant features as well as enhance cross-modality alignment. As opposed to the basic concept of feature concatenation, cross-attention dynamically assigns relevance to various components of features and generates more discriminative and informative representations, which eventually translate into improved classification.

6.3 Comparison with Baselines

Comparison with Baseline Methods

A thorough comparison, in order to assess the effectiveness of the proposed TransWalkCotNet model, is made with the methods of the baseline, which include CNN-based (ResNet18), Transformer-based (Swin Transformer), and classical machine learning methods. The findings show conclusively that the proposed model is always better than all the baselines in various evaluation measures.

1. Comparison with ResNet (CNN-based Model)

ResNet18 is a standard CNN example of a baseline that is based on convolutional operations as a primary means of extracting local spatial features. Though more complex architectures can be trained with residual connections, and its training stability is better, the model itself is inevitably constrained by the local receptive fields. It, therefore, does not take into account global contextual dependencies in various parts of the leaf image and does not assume relationships between various samples. Also, it has problems with situations that are characterized by subtle disease patterns or distributed in space. The experimental findings

mirror these limitations as ResNet has relatively lower performance in the accuracy, F1-score, and ROC-AUC.

2. Comparison with Swin Transformer (Visual Baseline)
The Swin Transformer is a better model as it improves the CNN-based models by using self-attention mechanisms that allow the model to extract global spatial relationships in images. This enables the model to be successful in capturing long-range interactions, recognizing the complicated visual patterns, and becoming more resistant to changes in the leaf structure and background. Nevertheless, even in its high-performance, Swin Transformer can only use visual data and lacks a specific implementation of inter-sample connections or structural resemblances between data points. This implies that it is unable to make use of the relational knowledge and might not utilize the underlying structure of the dataset to the full potential. This is the reason why despite the fact that Swin Transformer is superior to ResNet, it still loses against the suggested approach.

3. Advantage of the Proposed TransWalkCotNet
This will give the best results to the proposed TransWalkCotNet model that successfully combines the merits of both the visual and relational learning paradigms. This integrated architecture allows the model to learn more discriminative and informative features, intra-image features, inter-sample relationship and perform the classification significantly better in complex and unclear situations.

4. Result-based Interpretation of Performance
As shown in the comparison outcomes and ROC curves, the suggested model is constantly superior to the baseline methods in a variety of assessment measurements. It has a greater accuracy and F1-score, which means that it is more reliable with regards to classification, and the ROC curves show greater values of AUC that show better discriminative ability. Also, the cross-validation results demonstrate the consistency of the performance of the model with regard to various folds, which proves the stability and robustness of the model. These results illustrate the efficiency of the visual and relational learning applying in one framework.

6.4 Analysis of Traditional Machine Learning Methods
Besides deep learning baselines, a range of traditional machine learning algorithms, such as Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Gaussian Naive Bayes are also under consideration in order to compare them all within a comprehensive perspective. Although these approaches have some benefits, they do not work well in such complicated image classification problems like cotton disease detection.

1. Analysis of Patterns Performance
Conventional machine learning models tend to work well in the case of low-dimensional and simple feature representations. These models can produce moderately good performance when the characteristics are separated and have linear or moderately non-linear relationships. As an example, Logistic Regression is effective in situations where the decision boundary is roughly linear, the Random Forests can model moderate non-linear relations by using ensemble decision trees, and KNN is effective in classifying the samples according to local similarity in the feature space. These techniques are however very dependent on the quality of input features mostly obtained manually or used as shallow representations. These models can yield satisfactory outcomes in a situation where the visual patterns are simple, and the symptoms of the disease are distinct and clear.

2. Weaknesses in Management of Multifactorial Disease Characteristics

The visual features of cotton leaf diseases are normally complex and can include an abnormal texture, subtle color changes, overlapping patterns and changes with changing lighting and with background noises. Such situations are hard in traditional machine learning models since they cannot easily learn the hierarchical features representations, as well as representative the high dimensional feature interactions. Their performance is still worse when the features are not linearly separable. Indicatively, Logistic Regression cannot model the complex decision boundaries, the Random Forest can overfit or fail to generalize in seriously complicated feature spaces, KNN is afflicted with the curse of dimensionality, and Gaussian Naive Bayes makes strong independence assumptions that cannot be true of image data. These models therefore generally have lower accuracy and F1-scores than deep learning-based models.

3. Weakness in Spatial and Contextual Competency

The failure of traditional machine learning approaches to extract spatial and contextual information in images is one of the key limitations of the traditional machine learning techniques. They also cannot capture global context at various parts of the image unlike deep learning models because they do not maintain spatial connections between pixels, cannot learn local structures such as the edges, textures and shapes of objects, and cannot learn coherent patterns at multiple scales of the image. These models are mathematically defined on flattened feature vectors.

4. General functioning impairment

Combination of the abovementioned constraints means that traditional machine learning techniques cannot be fully used to represent cotton disease patterns and their complexity. The results of the experiment demonstrate this point, as these models are not only less accurate in the classification, the F1-scores and recall values are worse, and the ROC curves are drawn to show lower discriminative capability, which can be associated with more sophisticated methods.

Accepted ML Evaluation

Even though the traditional machine learning models are computationally efficient and can be used in simple tasks, they do not best fit complex image-based disease classification. They are not effective due to their inability to obtain spatial, hierarchical, and relational information. Conversely, the limitations are overcome in the proposed TransWalkCotNet model via deep visual feature extraction with Transformers, relational learning with graph-based techniques and DeepWalk embeddings, and feature fusion with cross-attention mechanisms. This broad methodology justifies the high-performance improvement as compared to the conventional machine learning methodologies.

6.5 Key Insight (IMPORTANT)

A combination of graph-based embeddings and transformer features is demonstrated to be an effective way to enhance the performance of classification by learning the relationship between samples. The most important observation made in this research is that proper classification of plant diseases involves not only good visual feature extraction, but also relational structure among samples. Conventional methods have also been more concerned with enhancing the visual representations but not paying significant attention to inter-sample dependencies.

1. Limit of the Single-View Learning

The majority of existing models, such as CNNs and Transformer-based ones, use a single-view paradigm of learning, in which each image is treated separately. Despite the fact that these models can extract powerful feature representations, they do not take relationships among the samples into consideration since they could provide good contextual information. This weakness lowers the capacity of the model to

distinguish well the visually similar disease categories, results in bad generalization when dealing with slight or confusing features and gives an imperfect depiction of the underlying data structure.

2. Significance of Relational Learning

The findings of this research point out the fact that the use of graph-based relational learning generates a substantial improvement in the classification performance. Using a similarity graph and DeepWalk embeddings, the model can effectively represent latent structural relationships among samples, cluster similar disease patterns, and can be more effective at decision boundaries in multifaceted feature spaces. This proves that the natural organization of the information is already valuable and should not be disregarded when performing the tasks of classifying.

3. Complementarity of Structural and Visual Features

One of the most important conclusions made in this work is that visual features and graph-based embeddings are complementary to each other. Transformer visual features are a representation of appearance and texture, and DeepWalk graph embeddings represent the relationship and similarity between samples. Differently, each of the representations gives biased information but when fused together via a cross-attention fusion mechanism, will create a holistic and more inclusive representation of the data. This integration allows more effective discrimination of related classes, is less susceptible to noise and variability, and overall is more effective at classification.

4. The Unified Hybrid Framework has several benefits

The proposed TransWalkCotNet introduces single hybrid learning framework, which combines Transformer-based global feature extraction, graph-based structure modeling, DeepWalk embedding learning, and cross-attention features fusion. This single design is effective in solving the main weakness of the past designs where these components are considered separately. As the outcomes of the experiment allow showing, such integration result in the significant and consistent model performance improvements.

5. Broader Implications

The conclusions made in this paper can be applied not only to cotton disease classification but can be applied in general to other fields like medical image analysis, biological data modeling, social network analysis, and recommendations systems. These results indicate that the deep learning along with graph-based relational learning is a promising research direction in the future, especially in scenarios that deal with complex and structured data.

6.6 Limitations of the Proposed Model

Although the work of the proposed TransWalkCotNet framework is better than its possible competitors, there are some limitations that should be considered to make a fair assessment of the model.

1. Computational Cost of DeepWalk

The main weakness of the suggested solution is that the DeepWalk algorithm is computationally complex. The DeepWalk algorithm works by creating several truncated random walks on every node in the graph and then training a skip-gram model to obtain node embeddings.

$$\mathcal{O}(|V| \cdot \gamma \cdot t) \quad (43)$$

This process can be estimated to have a time complexity of g where g is the number of walks that have to be made per node, and t is the length of a walk. The larger the size of the dataset, the greater are the number of nodes in the graph and the higher are the computational overhead, training time, and memory consumption. With this, DeepWalk is no longer effective with large datasets, especially on real-time processes.

2. Overhead of Graph Construction

The other weakness is because the construction of the KNN-based similarity graph involves calculation of pairwise similarities between feature vectors. This process can be described as computationally complex.

$$\mathcal{O}(N^2) \quad (44)$$

N is a symbol of the number of samples. Despite the fact that the cost can be lowered using approximate nearest neighbor methods, the graph construction process also adds some preprocessing time, memory overhead to store adjacency matrices and scalability is a problem when the size of the data is large. Moreover, the quality of the graph constructed much relies on the parameter selection especially the number of neighbors (k) and the similarity measure method. The inappropriate choice of these parameters can lead to inefficient graph structures, which can have a negative effect on the quality of downstream embeddings and the overall performance of the model.

7. Future Work

Despite the fact that the proposed TransWalkCotNet framework has shown good performance regarding the classification of cotton diseases, there are some potential avenues through which this study can be enhanced and expanded.

1. Integration with Graph Neural Networks (GNNs)

The graph-based representation learning on the existing framework is conducted with DeepWalk that learns the structural information by random walks. DeepWalk is however, a shallow embedding approach and it does not explicitly use node features in the learning process. To overcome this weakness, future investigations can consider the use of Graph Neural Networks (GNNs), including Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and GraphSAGE, which support end-to-end learning on the graph structure.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (45)$$

GNNs have a number of benefits, such as immediate access to node features in addition to graph structure, allow message transmission between nodes, and learn more expressive and adaptive representations. The use of GNN-based approaches would also help in improving the performance of classification by either substituting or complementing DeepWalk-based embeddings.

where:

- $\tilde{A}$is the adjacency matrix with self-loops
- $\tilde{D}$is the degree matrix
- $H^{(l)}$represents node features at layer l

2. Quantum Graph Learning Exploration

The use of quantum computing methods in graph-based learning is another potential research avenue. The basic principles of quantum superposition and entanglement have been used in quantum graph learning to learn complex relational data more effectively. To give an example, quantum walks can be used in place of classical random walks and are mathematically defined as

U is the unitary evolvement operator. The quantum methods could be more efficient than the classical methods in terms of exploring the graph structure, studying complex relationships in higher dimensional spaces and making large-scale problems less complex. Despite being at its initial phase, the combination of quantum graph learning and deep learning models would provide new research opportunities in intelligent agriculture and other interconnected fields.

3. Live Deployment and Live applications

Although the given model can be highly precise, it still needs to be optimized to be deployed in a real-life agricultural context. Future directions in this area can be geared towards lightweight and computationally

efficient model architectures, and minimizing computational overhead of edge devices, and deploying the model to mobile or embedded systems to operate at the field. The real-time deployment would allow the detection of diseases in real-time with smartphones cameras, support or aid farmers in precision agriculture, and intervene early in order to reduce crop loss. Moreover, interconnection with IoT-based smart farming can increase scalability, automation, and the ability to track in real-time.

8. CONCLUSION

This paper introduces a new hybrid architecture TransWalkCotNet, which is a combination of transformer-based visual learning and graph representation learning. The experiment findings validate that the suggested strategy performs better than both the deep learning baselines and the traditional machine learning techniques. The findings show the need to incorporate relational knowledge in order to classify agricultural diseases better.

In this paper, a new hybrid model, TransWalkCotNet, has been presented to classify cotton disease precisely and effectively. The model is efficient in combining Transformer-based visual features extraction with graph-based representation learning, which can overcome the shortcomings of the current methods that use either visual or structural data only. The Swin Transformer component allows extracting rich and global visual features that capture complex patterns of space that might exist in cotton leaf images. A cross-attention fusion mechanism is then used to combine these complementary representations, which enables the model to realize the positive feedback of using both visual and relational knowledge. Large-scale experiments on a dataset of cotton disease indicate that the proposed model is much more effective compared to deep learning baselines (including ResNet and Swin Transformer) and classical machine learning techniques. The model has a better performance on most measures of evaluation which are accuracy, precision, recall, F1-score and ROC-AUC, which means that its discriminatory performance and generalization performance are high.

The findings also indicate that the ability to include relational learning with the help of graph-based methods offers useful contextual information that improves the accuracy of classification, in particular, in complicated situations when the patterns of the diseases are hidden or overlapping. The cross-attention fusion is important in better aligning and integrating the heterogeneous features to come up with more informative representations. Although effective, the model presents more computational complexity because of the construction of graphs and generation of embeddings. Nevertheless, the future work in these areas can overcome these difficulties by providing more effective methods of graph learning and the optimization of their implementations. To sum up, the proposed framework of TransWalkCotNet is a potent and unified framework to classify plant diseases by filling the gap between deep learning based on visual representations and graph-based relational modeling. Besides pushing the state-of-the-art of cotton disease detection, one can see the comprehensive application of hybrid learning frameworks to challenging image classification problems in agriculture and other associated fields.

Author Contributions

H.S. and A.G. conceptualized and designed the study. M.S.H.T. and RAB formulated the methodology. The software implementation and experiments were done by R.S. Z.A.M. performed method validation. H.S. and H.S carried out the formal analysis. A.G. also helped in research and resources. R.S. and RAB did the data curation and the initial draft preparation. H.S. assisted in writing -review and editing. D.M.V and A.G. were in charge of visualization and supervision. The final version of the manuscript has been read by all the authors and approved by them.

Data Availability Statement

The datasets that have been used in the analysis in the present study can be found publicly in Kaggles, at https://www.kaggle.com/datasets/janmejaybhoi/cotton-disease-dataset, and https://www.kaggle.com/datasets/seroshkarim/cotton-leaf-disease-dataset[34] and https://data.mendeley.com/datasets/74jsdxtmx2/3 [47, 48]. The source code and the data used to present the results of this research can also be presented by the relevant author, in case of reasonable demand.

Funding
This research received no external funding.

Institutional Review Board Statement
Not applicable.

Informed Consent Statement
Not applicable.

Data Availability Statement
Suggested data availability statements are available on request.

Acknowledgments
The authors would like to thank Sind Agriculture University for their valuable support.

Conflicts of Interest
The authors declare no conflicts of interest

REFERENCES
1. Lu, J., Tan, L., & Jiang, H. (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. Agriculture, 11(8), 707.
2. Sunil, C.K., Jaidhar, C.D. & Patil, N. Systematic study on deep learning-based plant disease detection or classification. Artif Intell Rev 56, 14955–15052 (2023). https://doi.org/10.1007/s10462-023-10517-0
3. Shi, T., Liu, Y., Zheng, X., Hu, K., Huang, H., Liu, H., & Huang, H. (2023). Recent advances in plant disease severity assessment using convolutional neural networks. Scientific Reports, 13(1), 2336.
4. Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., ... & Ali, F. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in plant science, 14, 1158933.
5. Yi, H. C., You, Z. H., Huang, D. S., & Kwoh, C. K. (2022). Graph representation learning in bioinformatics: trends, methods and applications. Briefings in Bioinformatics, 23(1), bbab340.
6. Ning, Q., Zhao, Y., Gao, J., Chen, C., Li, X., Li, T., & Yin, M. (2023). AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA–disease associations identification. Briefings in Bioinformatics, 24(2), bbad094.
7. Dong, W., Moses, C., & Li, K. (2011, March). Efficient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th international conference on World wide web (pp. 577-586).
8. Shafi, H., Ghulam, A., Talpur, M. S. H., & Sikander, R. (2026). Heterogeneous Network Framework for Predicting Novel Disease–Plant Associations Using Random Walk with Restart (RWR). AgriEngineering, 8(3), 113.
9. Dhanabal, S., & Chandramathi, S. J. I. J. C. A. (2011). A review of various k-nearest neighbor query processing techniques. International Journal of Computer Applications, 31(7), 14-22.

10. Liu, X., Liu, R., Li, F., & Cao, Q. (2012, November). Graph-based dimensionality reduction for KNN-based image annotation. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 1253-1256). IEEE.
11. Shafi, H., Ghulam, A., Talpur, S. H., Sikander, R., Ali, A., Jabeen, N., ... & Iskandar, Y. (2025). A Comprehensive Review of Complex Network Methods for Cotton Plant Disease Detection. Journal of Information Communication Technologies and Robotic Applications, 16(1).
12. Balafas, V., Karantoumanis, E., Louta, M., & Ploskas, N. (2023). Machine learning and deep learning for plant disease classification and detection. IEEE Access, 11, 114352-114377.
13. Elngar, A. A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., & Fawzy, N. (2021). Image classification based on CNN: a survey. Journal of Cybersecurity and Information Management, 6(1), 18-50.
14. Karthik, R., Ajay, A., Jhalani, A., Ballari, K., & K, S. (2025). An explainable deep learning model for diabetic foot ulcer classification using swin transformer and efficient multi-scale attention-driven network. Scientific Reports, 15(1), 4057.
15. Bossoun, K. K. H., & Ying, X. (2024, December). A graph neural network approach for early plant disease detection. In International Conference on Data Mining and Big Data (pp. 254-265). Singapore: Springer Nature Singapore.
16. Keikha, M. M., Rahgozar, M., & Asadpour, M. (2018). Community aware random walk for network embedding. Knowledge-Based Systems, 148, 47-54.
17. Vishnoi, V. K., Kumar, K., & Kumar, B. (2022). A comprehensive study of feature extraction techniques for plant leaf disease detection. Multimedia Tools and Applications, 81(1), 367-419.
18. Chen, Z., Wu, R., Lin, Y., Li, C., Chen, S., Yuan, Z., ... & Zou, X. (2022). Plant disease recognition model based on improved YOLOv5. Agronomy, 12(2), 365.
19. Jadhav, S. B., Udupi, V. R., & Patil, S. B. (2021). Identification of plant diseases using convolutional neural networks. International Journal of Information Technology, 13(6), 2461-2470.
20. Mascarenhas, S., & Agarwal, M. (2021, November). A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON) (Vol. 1, pp. 96-99). IEEE.
21. Liang, J. (2020, September). Image classification based on RESNET. In Journal of Physics: Conference Series (Vol. 1634, No. 1, p. 012110). IOP Publishing.
22. Shi, Q., Tang, X., Yang, T., Liu, R., & Zhang, L. (2021). Hyperspectral image denoising using a 3-D attention denoising network. IEEE Transactions on Geoscience and Remote Sensing, 59(12), 10348-10363.
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
24. Pan, X., Ye, T., Xia, Z., Song, S., & Huang, G. (2023). Slide-transformer: Hierarchical vision transformer with local self-attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2082-2091).
25. Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE international conference on computer vision (pp. 1449-1457).
26. Awan, K., Asadullah Shaikh, M. R., Alsulami, M., Alqhtani, S., & AlYami, S. (2026). Dynamic graph-enhanced deep network for robust infected fruit detection through adaptive CNN-GNN feature fusion. Journal of King Saud University Computer and Information Sciences.

27. Dong, W., Moses, C., & Li, K. (2011, March). Efficient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th international conference on World wide web (pp. 577-586).
28. Chanpuriya, S., Musco, C., Sotiropoulos, K., & Tsourakakis, C. (2021, July). Deepwalking backwards: from embeddings back to graphs. In International conference on machine learning (pp. 1473-1483). PMLR.
29. Shafi, H., Ghulam, A., Talpur, M. S. H., & Sikander, R. (2026). Transformer-Based Cotton Plant Disease Detecting: Comparative Study to Deep Learning and Machine Learning, Gongcheng Kexue Yu Jishu/Advanced Engineering Science, Journal ID : AES-23-03-2026-934, Volume - 58, Issue - 02.
30. Rehman, A., Naz, S., Razzak, M. I., & Hameed, I. A. (2019). Automatic visual features for writer identification: a deep learning approach. IEEE access, 7, 17149-17157.
31. Zhong, M., Wei, L., & Mo, H. (2025). Cotton pest and disease diagnosis via YOLOv11-based deep learning and knowledge graphs: a real-time voice-enabled edge solution. Frontiers in Plant Science, 16, 1671755.
32. Kumar, E. P., Malathi, S., & Yadav, C. S. B. (2026). Progressive Regional Spatial Mechanism With Convolutional Neural Network For Cotton Leaf Disease Classification. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES, 569-582.
33. Chohan, S., Perveen, R., Abid, M., Tahir, M. N., & Sajid, M. (2020). Cotton diseases and their management. In Cotton production and uses: agronomy, crop protection, and postharvest technologies (pp. 239-270). Singapore: Springer Singapore.
34. Rai, C.K., Pahuja, R. An ensemble transfer learning-based deep convolution neural network for the detection and classification of diseased cotton leaves and plants. Multimed Tools Appl 83, 83991–84024 (2024). https://doi.org/10.1007/s11042-024-18963-w
35. Ferdous, G. J., Sathi, K. A., Hossain, M. A., & Dewan, M. A. A. (2024). SPT-Swin: A shifted patch tokenization Swin transformer for image classification. IEEE Access, 12, 117617-117626.
36. Kang, S. (2021). K-nearest neighbor learning with graph neural networks. Mathematics, 9(8), 830.
37. Dad, I., He, J., & Baloch, Z. (2024). Graph-Based Analysis of Histopathological Images for Lung Cancer Classification Using GLCM Features and DeepWalk Embeddings.
38. Chanpuriya, S., Musco, C., Sotiropoulos, K., & Tsourakakis, C. (2021, July). Deepwalking backwards: from embeddings back to graphs. In International conference on machine learning (pp. 1473-1483). PMLR.
39. Zheng, J., Liu, H., Feng, Y., Xu, J., & Zhao, L. (2023). CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation. Computer Methods and Programs in Biomedicine, 229, 107307.
40. Sarangdhar, A. A., & Pawar, V. R. (2017, April). Machine learning regression technique for cotton leaf disease detection and controlling using IoT. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 2, pp. 449-454). IEEE.
41. Mitra, A., Beegum, S., Fleisher, D., Reddy, V. R., Sun, W., Ray, C., ... & Malakar, A. (2023). Cotton yield prediction using random forest. arXiv preprint arXiv:2312.02299.
42. Kursun, R., & Koklu, M. (2024). Optimized Classification of Cotton Leaf Diseases Using Machine Learning and Feature Selection Techniques. Agri-Intelligence. Çizgi Kitabevi Publishing.[Google Scholar].
43. Atoyebi, T. O., Olanrewaju, R. F., Blamah, N. V., & Uwazie, E. C. (2024, April). Comparison of multinomial naive Bayes (MNB), Gaussian naive Bayes (GNB) and random forest (RF) algorithm in malaria disease diagnosis. In 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG) (pp. 1-6). IEEE.
44. Singh, A., Rakhra, M., Raj, R., Potharaju, S., KANTIPUDI, M. P., Gowroju, S., & NS, P. K. (2026). Validated Swin Transformer-Based Deep Learning Pipeline with Cross-Validation and McNemar's

Test for Multi-Class Hemorrhage Classification in Traumatic Brain Injury CT Scans. Journal of Artificial Intelligence and Technology, 6, 224-235.
45. Kang, Z., Xiao, E., Li, Z., & Wang, L. (2024). Deep learning based on ResNet-18 for classification of prostate imaging-reporting and data system category 3 lesions. Academic Radiology, 31(6), 2412-2423.
46. Jin, X., Zhou, J., Rao, Y., Zhang, X., Zhang, W., Ba, W., ... & Zhang, T. (2023). An innovative approach for integrating two-dimensional conversion of Vis-NIR spectra with the Swin Transformer model to leverage deep learning for predicting soil properties. Geoderma, 436, 116555.
47. Agarwal, Meet (2025), "Cotton Leaf Disease Dataset with Severity Levels", Mendeley Data, V1, doi: 10.17632/74jsdxtmx2.1
48. Shingne, Shraddha; Thakur, Parth; Jangid, Abhishek; Buchade, Punam; Jain, Ayush; Agarwal, Meet (2025), "Cotton Leaf Disease Dataset with Severity Levels", Mendeley Data, V3, doi: 10.17632/74jsdxtmx2.3