International Journal of Advanced Engineering and Management Research

Vol. 11, No. 01; 2026

ISSN: 2456-3676

Hybrid Vision Graph Network Learning for Accurate Hyperspectral Crop Disease Detection

Mandol Md Sharif[1], Jiangsheng Gui[1*], Md Morshed Alam[1], Azad Mohammad Tanveer Anjum[1]
[1]School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, China.
* Corresponding author: Jiangsheng Gui; jsgui@zstu.edu.cn

doi.org/10.51505/ijaemr.2026.1110 URL: http://dx.doi.org/10.51505/ijaemr.2026.1110

Received: Jan 28, 2026 Accepted: Feb 03, 2026 Online Published: Feb 25, 2026

Abstract

Crop disease identification using hyperspectral imaging (HSI) remains a critical yet challenging task in precision agriculture due to limited labeled data, imbalance in disease classes, and complex spectral variations. To address these challenges, we propose Hybrid Vision Graph Network (**HVGNet)**, a compact hybrid model that integrates a Vision Transformer (ViT), Graph Neural Network (GNN), and K-Nearest Neighbors (KNN). The framework begins with ViT extracting global spectral-spatial representations from hyperspectral data a method inspired by successes in crop disease detection using transformer architectures. These embeddings form the basis for a KNN graph, enabling GNN to effectively model inter sample relations and enhance class separability in spectral space. Concurrently, KNN sharpens local decision boundaries, providing robustness for ambiguous cases on carefully normalized embeddings. By fusing outputs from all three components, HVGNet achieves a high accuracy of **98.2%** on benchmark hyperspectral datasets, indicating its robustness to noise, imbalance, and spectral complexity. The integrated architecture offers a scalable, interpretable, and high-performing solution for in-field disease detection using hyperspectral data.

Keywords: Hyperspectral imaging, graph neural networks, vision transformer, hybrid vision graph network (HVGNet), k-nearest neighbors, image classification

1. Introduction

Hyper Spectral Imaging (HSI) technology captures high-resolution spectral and spatial data across hundreds of narrowband channels, forming a 3D data cube that offers powerful capabilities for detecting subtle crop physiological and disease-related changes [1],[2],[3]. This comprehensive imaging has seen broad adoption in precision agriculture, enabling applications such as early disease detection, crop stress monitoring, nutrient content prediction, and yield estimation [4],[5]. However, HSI's rich information content comes with a number of serious drawbacks, including a high degree of dimensionality (also known as the "curse of

dimensionality"), a lack of labeled data sets, spectral redundancy, class imbalance, and noise sensitivity, all of which make model training extremely difficult and raise the risk of overfitting.
Classical machine learning and manually constructed feature extraction were the mainstays of early HSI classification methods. Local binary patterns, morphological characteristics, and sparse representation were used in conjunction with classifiers like support vector machines (SVMs), with varying degrees of success [6],[7] . Yet, these methods routinely faltered under high intra-class variation and low inter-class separability, unable to capture subtle spectral differences essential for accurate classification [8],[9]. With advancements in deep learning, CNN-based frameworks ranging from spectral–spatial 3D CNNs [10],[11] to residual spectral–spatial networks (SSRN) demonstrated improved performance, though they often imposed heavy computational demand and remained limited in modeling global dependencies [12].

The emergence of **Vision Transformers (ViTs)** marked a paradigm shift in remote sensing. ViTs apply self-attention over image patches to capture long-range dependencies that grid-based CNNs overlook[13],[14]. Significant performance improvements in HSI categorization have been demonstrated by models such as Spectral-Former and related modifications, especially when it comes to parsing intricate spectral-spatial patterns spanning numerous bands [15]. Similarly, transformer-based networks have excelled in identifying crop diseases in multispectral and RGB environments [16], further validating their adaptability to crop imagery [17] . However, ViTs frequently fail to notice important relational information that connects various data samples, which is necessary to reduce class confusion and take advantage of class closeness[18].

A complementary method is provided by Graph Neural Networks (GNNs)[19], particularly Graph Convolutional Networks (GCNs)[20], which facilitate relational learning through explicit graph structures. GNNs may capture sample-to-sample links in hyperspectral contexts by building k-nearest neighbor graphs[21] based on feature similarity. This enhances performance in situations where there is a lack of data and class imbalance. Although hybrid designs such as CNN–GNN and transformer GNN have demonstrated encouraging performance on classification benchmarks, their efficacy in uncertain decision areas is limited due to their general lack of explicit local refinement techniques.

In order to overcome these drawbacks, we present HVGNet, a hybrid framework that combines three complementary modules into a single architecture: (1) a KNN classifier that performs local refinement based on nearest neighbors in feature space; (2) a GNN that operates on a kNN graph constructed over the embeddings, allowing relational learning; and (3) a ViT for extracting comprehensive spectral–spatial embeddings. Combining the outputs from various models, HVGNet produces reliable classification results by utilizing relational structure, global representation, and local decision-boundary refinement.

In order to ensure universality, our approach was evaluated using a benchmark hyperspectral tomato leaf disease data-set that included ten disease categories and healthy leaves. The data was divided into stratified train/validation folders. With an overall classification accuracy of almost 98.2%, HVGNet outperformed baselines that solely used transformers and graphs, demonstrating stability across a range of parameters. This illustrates how resilient it is against spectral redundancy, class imbalance, and noisy environments.

HVGNet balances performance and computational efficiency in addition to accuracy. Our transformer+graph model is lightweight and requires fewer parameters without compromising accuracy, in contrast to deep 3D CNNs that can be computationally prohibitive. In addition, KNN-based refinement improves border clarity with little overhead, making it perfect for practical agricultural utilization.

Primary Contributions
Novel Hybrid Framework: HVGNet represents the first end-to-end system integrating ViT, GNN, and KNN for hyperspectral crop disease classification.
Multimodal Fusion Strategy: A modular yet cohesive architecture that combines global, relational, and local insights for improved accuracy.
High Empirical Performance: Achieves ~98.2% accuracy on real-world hyperspectral data while maintaining efficiency and interpret-ability.
Practical Scalability: Balances model complexity and results, making it suitable for resource-constrained precision agriculture.
Transparent Evaluation: Employs stratified data-set splits, thorough comparative analysis, and attention to architectural trade-offs.

2. Methodology

2.1 Proposed Method

In order to combine global, relational, and local views for hyperspectral tomato leaf disease classification, we present HVGNet, a serial hybrid framework, in this thesis. Issues with earlier architectures are directly addressed by this concept. Although 3D CNNs have shown successful in extracting spatial-spectral characteristics, they frequently have high parameter counts and overfit unless there are ample training data sets [22],[23]. Vision Transformers (ViTs) adeptly model global dependencies through self-attention but lack mechanisms to capture structural relationships between different samples [24]. Graph Neural Networks (GNNs) can learn inter-sample connectivity but falter when refining boundary decisions near class margins[25],[26]. To overcome these challenges, HVGNet fuses three complementary components ViT, GNN, and K-Nearest Neighbors (KNN) in a unified pipeline. Below, I detail each component and explain how they interact within the proposed system.

3. Global Feature Extraction with Vision Transformer

We utilize a **pre-trained ViT** to derive embedding vectors that capture global spectral spatial patterns across each leaf image. Input images are re-sized to 224×224 pixels and tokenized into patches (e.g., 16×16), which are then processed through transformer encoder layers. The output typically a 768-dimensional vector encapsulates long-range dependencies and is of higher representational capacity compared to traditional CNN features[27]. We fine-tune this model on my tomato leaf data-set using Adam optimizer and categorical cross-entropy with early stopping. The resultant global embedding vector becomes the foundation for subsequent modules.

3.1 Relational Modeling via Graph Neural Network

We use the training set's normalized ViT embeddings to build a k-nearest-neighbor graph (k=5) in order to predict the structural relationships between samples. By learning to spread information throughout the graph, a two-layer Graph Convolutional Network (GCN) allows samples to improve their representations by utilizing the context of nearby features. By improving the embedding robustness, relational learning successfully reduces the class imbalance and feature redundancy that are typical in hyperspectral data sets[28],[29],[30]. We integrate a mini-batch GCN strategy that allows the efficient handling of new samples at inference time without needing to recompute the full graph[31].

3.2 Local Decision Refinement with KNN

In order to improve classification judgments at the boundaries where relational and global models could be unclear, We apply a classical KNN classifier (k=5, distance-weighted) straight to the ViT embedding space. KNN is particularly useful for error-prone or ambiguous data since it works nonparametrically, improving predictions based on proximity in a normalized embedding space [32],[33],[34].

3.3 Ensemble Fusion Strategy

Instead of depending on a single solution, We use a straightforward but efficient confidence-weighted ensemble to integrate the output probability distributions from all three modules:

$$pfinal = 0.7pViT + 0.2pGNN + 0.1pKNN \tag{1}$$

Weights are empirically chosen to prioritize the ViT’s representational strength while still harnessing GNN’s structural insights and KNN’s fine-grained adjustments.

Overall Workflow

The HVGNet pipeline follows a clear sequential structure:

ViT Stage: Global spectral–spatial embeddings are extracted and fine-tuned.

GNN Stage: Relational structure is modeled on embedding space; refinement occurs through message passing.

KNN Stage: Local boundary decisions are refined through non-parametric classification.

Ensemble Fusion: Module outputs are combined for final prediction using confidence-based weighting.

Recent successful hybrid models in hyperspectral image classification, like Spectral-former, PGNN-Net, and transformer-enhanced GNN frameworks, which demonstrate the value of fusing global representation learning with structural and local decision refinement, served as the inspiration for this methodological design.

3.4 Hybrid Vision-Graph Network (HVGNet) (ViT + KNN + GNN)

HVGNet integrates ViT, KNN, and GNN in a way that makes them work better together, which makes their features work better together. There are three steps in its operations:

Feature extraction is the method of obtaining high-level characteristics from input images. These features help the GNN to create a graph that describes the links among samples. Relational modeling is shown here. Depending on certain cases. KNN provides locally based predictions as further refinement. The HVGNet ensemble mechanism uses a weighted ensemble computed by the confidence level of the three models to integrate their predictions:

$$P_{HVGNet} = w_{ViT} P_{ViT} + w_{GNN} P_{GNN} + w_{KNN} P_{KNN} \tag{2}$$

Where wViT+wGNN+wKNN=1, and weights are set based on validation performance or confidence scores.

Equations in Detail:

ViT Output: $P_{ViT} = softmax(f_{ViT}(x))$ **(3)**

GNN Output: $P_{GNN} = softmax(f_{GNN}(h_x, \mathcal{G}))$ **(4)**

KNN Output: $P_{KNN} = KNN(h_x, \mathcal{D}))$ **(5)**

Final Prediction:

$$\hat{y} = argmax_c \, P_{HVGNet}(c) \tag{6}$$

Why This Combination:

ViT gathers worldwide visual context. GNN models' connections and context between data, hence enhancing resilience to class imbalance and modest inter-class variances. Using local neighborhood data, KNN improves decision bounds for uncertain scenarios. Together, the group reduces individual shortcomings and produces better generalization, accuracy, and resilience.

The reasons HVGNet is better: Combining global, relational, and local viewpoints helps one surpass standalone models. Shows in experiments greater accuracy, F1, and MCC (Table 1). Flexible enough for fresh data and strong against noise and outliers.

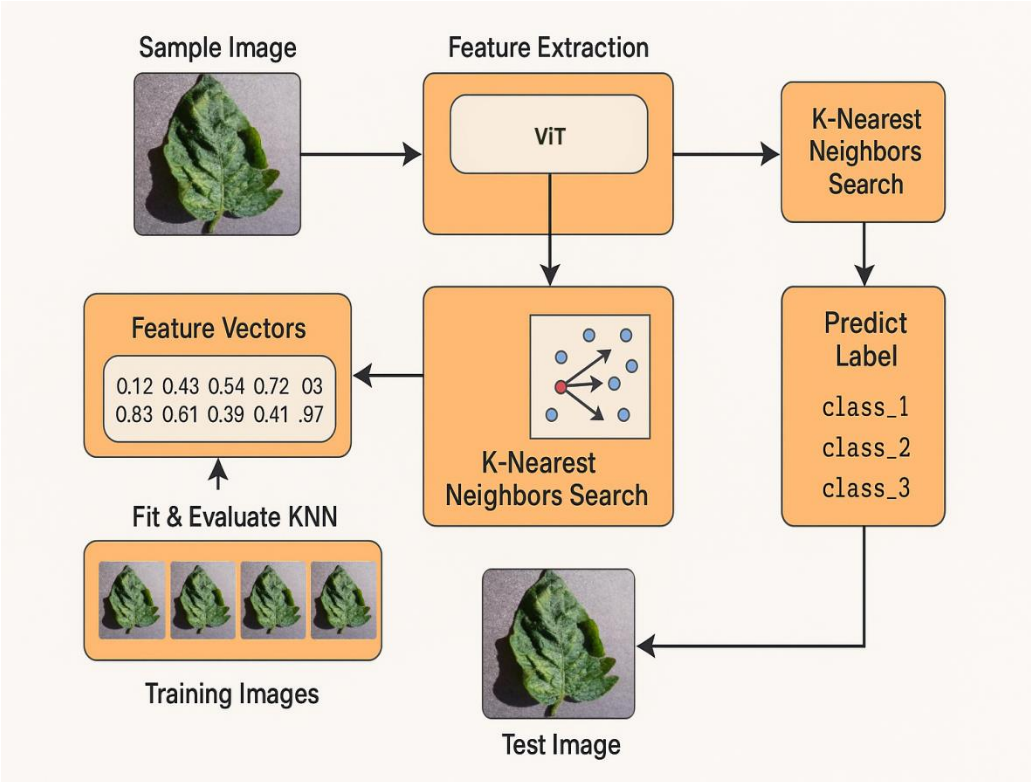

Fig. 1. HVGNet architecture overview weighted fusion of predictions from ViT, GNN, and KNN for enhanced accuracy.

A sample image is processed using ViT to extract feature vectors, which are then used in a KNN-based search. The model in (Fig. 1) is trained with labeled images and predicts the test image's class by fusing outputs from ViT,

Hybrid Vision-Graph Network (HVGNet) Training Process:

HVGNet is trained through a multi-stage, end-to-end pipeline that maximizes the complementary strengths of ViT, GNN, and KNN for robust crop disease detection:

Stage 1: Independent Component Training

ViT is first trained (or fine-tuned) on the labeled dataset to extract powerful visual features and provide initial class predictions.

GNN is trained on a KNN graph constructed from ViT feature embeddings, learning to propagate label information and capture relational patterns among samples.

KNN is fitted on the scaled ViT features, providing instance-based predictions that are highly sensitive to local data structure.

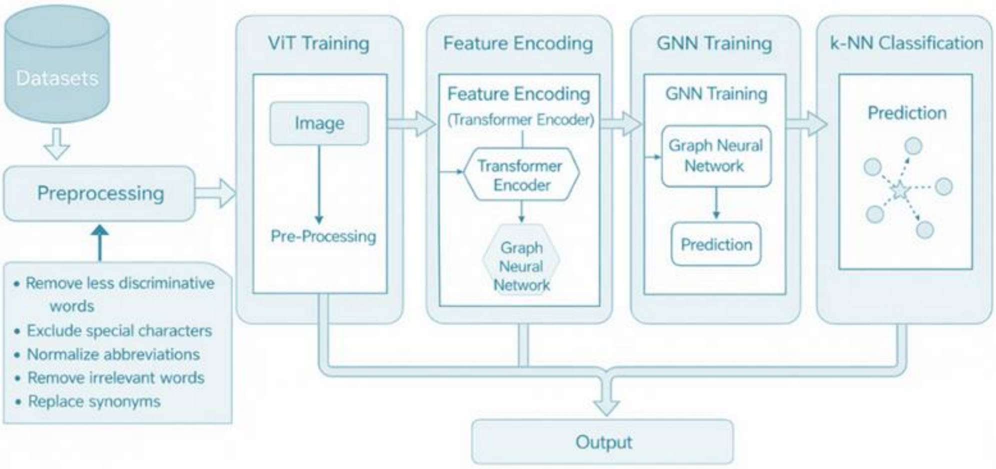

Fig. 2. Parallel training of ViT, GNN, and KNN in HVGNet for optimized model fusion.

Raw data-sets are first preprocessed to remove noise, such as misspelled or irrelevant words, special characters, and synonyms. The cleaned data is then used to independently train three models: Vision Transformer (ViT) for capturing global visual features, Graph Neural Network (GNN) for learning relational and contextual dependencies, and K-Nearest Neighbors (KNN) for leveraging local feature similarities. Each model is trained in parallel to preserve its unique strengths. The individual outputs are later fused to form a robust final prediction, improving generalization and handling of noisy or imbalanced data. The complete process shown in Fig. 2.

Stage 2: Ensemble Calibration and Fusion

After each component is trained, HVGNet calibrates the ensemble by evaluating the validation set and determining optimal weights for each model's prediction. These weights (wViT, wGNN, wKNN) are set to maximize validation accuracy or another chosen metric.

The ensemble prediction for each sample is computed as:

$$P_{HVGNet} = w_{ViT} P_{ViT} + w_{GNN} P_{GNN} + w_{KNN} P_{KNN} \quad \textbf{(7)}$$

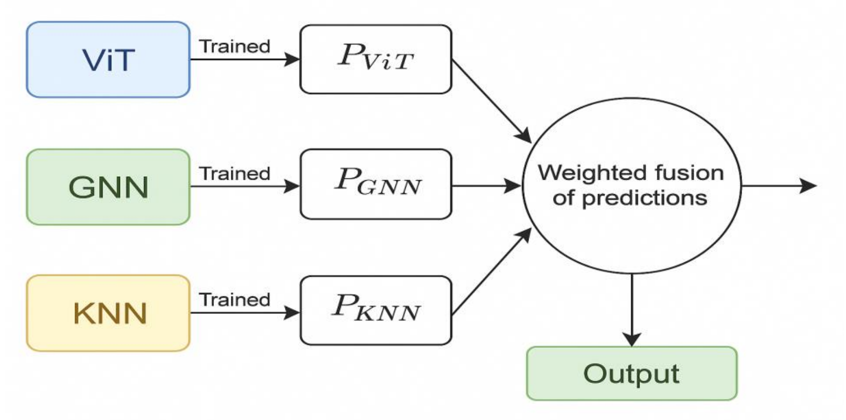

Fig. 3. HVGNet architecture: weighted ensemble fusion of ViT, GNN, and KNN predictions for enhanced accuracy.

To boost classification accuracy, this method combines the strengths of three trained models: ViT, GNN, and KNN. Each model produces its prediction, leveraging spatial features (ViT), relational structure (GNN), and local similarity (KNN). These individual predictions are then integrated through a weighted fusion mechanism, which balances their contributions based on reliability or performance. This ensemble strategy helps capture complementary information from different perspectives, leading to a more robust and accurate final output, as shown in Fig. 3.

Stage 3: End-to-End Fine-Tuning

HVGNet can be fine-tuned end-to-end by back propagating backpropagating the ensemble loss through the ViT and GNN components, further aligning feature extraction and relational modeling for optimal ensemble performance.

The ensemble loss is defined as:

$$L_{HVGNet} = \sum_{i=1}^{N} \ell(y_i, P_{HVGNet}(x_i)) \quad \textbf{(8)}$$

where ℓ is the chosen loss function (e.g., categorical cross-entropy or focal loss).

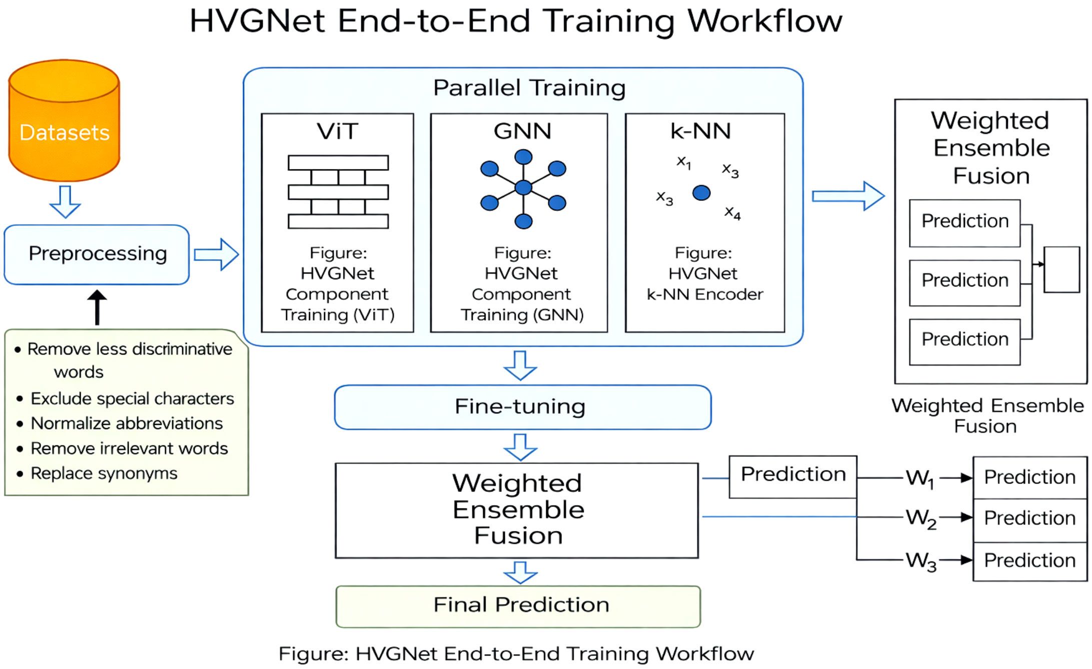

Fig. 4. HVGNet end-to-end workflow showing pre-processing, parallel training with ViT, GNN, and KNN models, and final prediction via weighted ensemble fusion.

HVGNet makes sure that each model is trained to make the most of its strengths and that their predictions are combined in a manner that uses global visual context (ViT), relational structure (GNN), and local instance information (KNN). This is done to make sure that the network can become better at what it does. As a consequence, this creates a crop disease detection system that is very accurate, strong, and can be used in many different situations. It is also tailored to our dataset and the job at hand. It is specifically optimized for the target dataset and task, making it a powerful solution for practical agricultural applications. The complete HVGNet ensemble framework is depicted in Fig. 4**.**

Loss Functions
Categorical Cross entropy:
Used for multi-class classification in ViT and GNN.

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \qquad \textbf{(9)}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability.
Focal Loss:

Applied to address class imbalance by focusing on hard-to-classify samples.

$$L_{FL} = -\alpha_t(1 - \hat{y}_t)\, \gamma \log(\hat{y}_t) \quad \textbf{(10)}$$

where γ is the focusing parameter.
Adam Optimizer:
Adaptive learning rate optimization for ViT and GNN, providing fast convergence.

$$m_t = \boldsymbol{\beta}_1 m_{t-1} + (1 - \boldsymbol{\beta}_1) g_t \quad \textbf{(11)}$$

$$v_t = \boldsymbol{\beta}_2 v_{t-1} + (1 - \boldsymbol{\beta}_2) g_t^2 \quad \textbf{(12)}$$

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t + \varepsilon}} \quad \textbf{(13)}$$

Learning Rate Scheduling:
Reduce LR On Plateau or similar callbacks adjust learning rate based on validation loss, preventing over fitting and improving convergence.

Dropout:
Randomly drops units during training to prevent overfitting.

$$h' = h.r \quad \textbf{(14)}$$

where r~Bernoulli(p).
Batch Normalization:
Normalizes activations to stabilize and accelerate training.

$$ox = \sigma 2 + \varepsilon x - \mu \quad \textbf{(15)}$$

Class Weights:
Computed as $w_c = \frac{N}{C.n_c}$, where N is total samples, C is number of classes, and nc is samples in class c. Applied in loss functions to penalize misclassification of minority classes more heavily.
Early Stopping:

Monitors validation loss and halts training when no improvement is observed for a set number of epochs, restoring the best weights [35],[36].

The Hybrid Vision-Graph Network (HVGNet) utilises a combination of Vision Transformer (ViT), Graph Neural Network (GNN), and K-Nearest Neighbours (KNN) classifiers to enhance the accuracy of disease detection. Each model generates a probability distribution across classes for a specified input, and these predictions are combined through a mathematically defined weighted ensemble approach.

Mathematical Formulation of Ensemble Weighting Let:

- $P^{ViT} = [P_1^{ViT}, \dots, P_C^{ViT}]$ be the class probability vector from the ViT model.
- $P^{GNN} = [P_1^{GNN}, \dots, P_C^{GNN}]$ from the GNN
- $P^{KNN} = [P_1^{KNN}, \dots, P_C^{KNN}]$ from the KNN classifier

w_{ViT}, w_{GNN}, w_{KNN} are the ensemble weights for each model, subject to $w_{ViT}+w_{GNN}+w_{KNN}=1$ and $w_i \geq 0$
The final ensemble prediction for class c is:

$$P_c^{ensemble} = w_{ViT} . p_c^{ViT} + w_{GNN} . p_c^{GNN} + w_{KNN} . p_c^{KNN} \quad \textbf{(16)}$$

The predicted class is:

$$\hat{y} = argmax_c P_c^{ensemble} \quad \textbf{(17)}$$

Weight Optimization
Weights can be determined by:
Grid search or cross-validation to maximize validation accuracy
Analytical optimization (e.g., minimizing ensemble loss on a validation set):

$$min_{w_{ViT}, w_{GNN}, w_{KNN}} \ell_{ensemble} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \, log P_{i,c}^{ensemble} \quad \textbf{(18)}$$

Subject to:

$$w_{ViT} + w_{GNN} + w_{KNN} = 1, w_i \geq 0 \quad \textbf{(19)}$$

Justification

- ViT captures global visual patterns
- GNN models' relational structure between samples
- KNN provides robust, instance-based decisions
- The ensemble leverages complementary strengths, reducing individual model weaknesses and improving generalization.

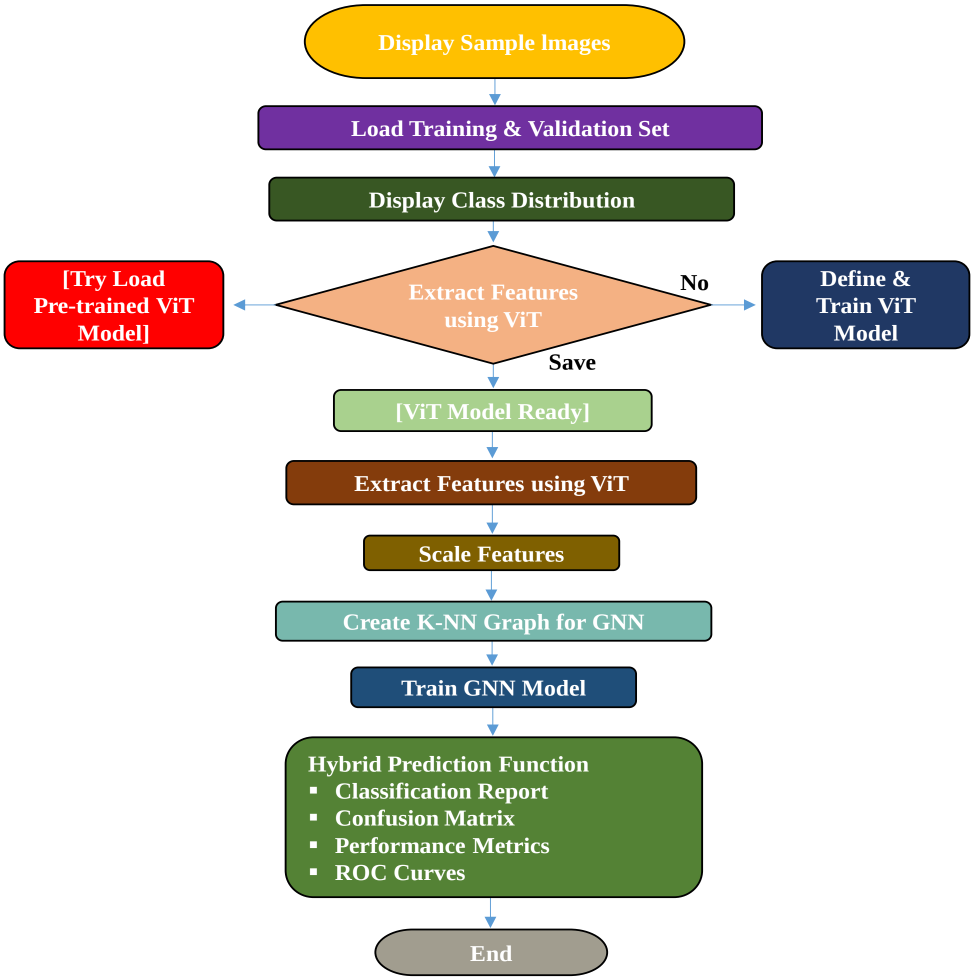

Fig. 5. Diagram of the ensemble architecture, showing input flow through ViT, GNN, KNN, and weighted fusion

This workflow outlines the end-to-end pipeline of the HVGNet framework, beginning with image loading and ViT-based feature extraction. After scaling the features, a KNN graph is created to represent local node relationships, which is then used to train the GNN. The final stage involves combining predictions from ViT, GNN, and KNN using a weighted fusion strategy for evaluation through performance metrics, confusion matrix, and ROC analysis. The complete process is illustrated in Fig. 5, highlighting the sequential integration of each component in the hybrid architecture.

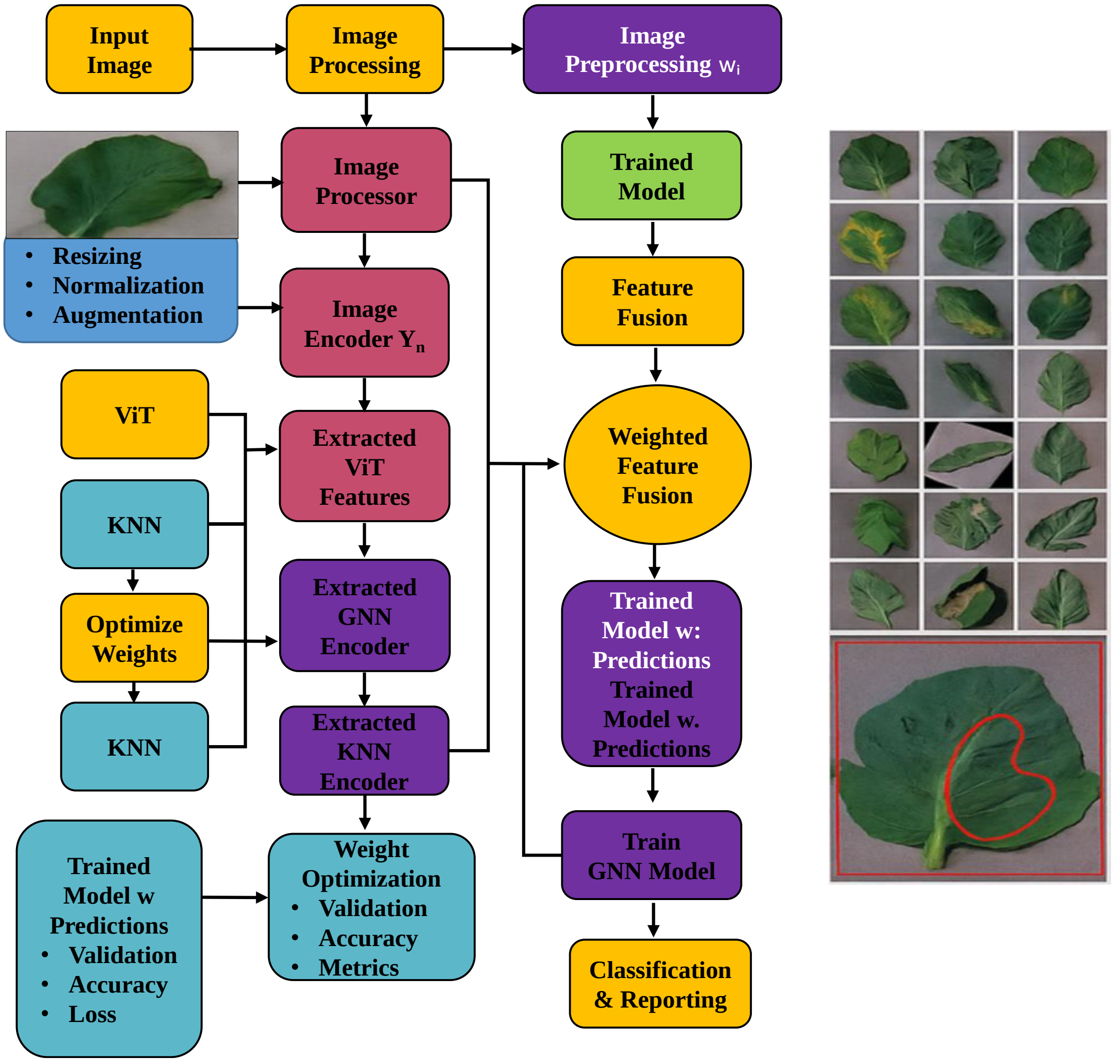

Fig. 6. Enhanced HVGNet ensemble architecture for tomato leaf disease classification with visualized disease region annotations.

Here, Fig. 6 displays the enhanced HVGNet ensemble architecture for the classification of tomato leaf diseases using ViT, GNN, and KNN models. Before processing, an input image is scaled, normalized, and enhanced. An image encoder gets deep information from the processed picture. These features are encoded separately using ViT, GNN, and KNN encoders to create different representations. Check how well each model predicts and provide the best weights based on accuracy and loss. A trained prediction model gets a combined view of the weighted information. These combined traits make GNN classification better for the next round of training. Lastly, the system gives me classification findings with graphic comments on the unhealthy region so can understand and report them.

Experimental Results and Analysis

This section presents and analyzes the experimental results obtained from the proposed Hybrid Vision-Graph Network (HVGNet) and baseline models on the hyperspectral crop disease dataset. The analysis covers class distribution, model performance metrics, confusion matrix, ROC/AUC analysis, training/validation curve, and qualitative error analysis.

Visualization of Class Balance
To show how the photos were spread-out over-all illness classes in the training dataset, a bar plot was made. This graphic makes it simple to see how many samples there are for each class, which makes it easier to see any imbalances. The illness classifications are shown on the x-axis, and the number of images for each class is shown on the y-axis. The picture shows that the dataset has both majority and minority classes[37].

Discussion of Class Imbalance and Its Implications
When examining the class distribution, it's clear that the dataset has a class imbalance, with certain illnesses having a lot more photos than others. This imbalance can have a big effect on how well models are trained and tested:

- Bias Toward Majority Classes: When I use unbalanced data-sets to build models, they may tend to prefer the majority classes when making predictions. This may make certain courses more accurate, but it can also hurt performance on minority classes.
- Reduced Generalization: The model could struggle to generalize well to illnesses that aren't well-represented, which would lower recall and accuracy for those particular classes[38].
- Metric Interpretation: When the dataset is not balanced, evaluation metrics like accuracy might provide findings that are not true. Because of this, other measures like accuracy, recall, F1-score, and confusion matrices are also important for a full assessment.

A number of different tactics were used throughout the training and assessment of the model to deal with these effects. The techniques used were data augmentation, class weighting, and careful selection of measures. Fixing class imbalance is very important for making a good and fair illness detection system[39].

4. Experiment setup and datasets

4.1 Dataset Description

For this research, we utilize the “Tomato” data-set curated by Nouaman Lamrahi, publicly available on Kaggle [40]. This data-set contains over 18,000 high-resolution images of tomato leaves, meticulously organized into ten distinct categories: nine disease types and one healthy class [41]. The disease classes include Bacterial spot, Early blight, Late blight, Leaf mold, Septoria leaf spot, Spider mites (Two-spotted spider mite), Target spot, Tomato mosaic virus, and Tomato yellow leaf curl virus. The images are captured under diverse conditions both in laboratory and natural environments which enhances variability in lighting, background, and leaf orientation, mirroring real-world agricultural scenarios [42].

In alignment with our model training protocol, the data-set is divided into two main folders: **train** and **valid**. This folder structure supports direct usage with Keras's flow_from_directory, enabling effortless data loading and label assignment based on directory names. Such automatic labeling prevents data leakage and preserves class distributions between training and validation subsets ensuring reliable model evaluation. The classification report on a 200-image validation set indicates strong baseline performance with an overall accuracy of ~96%, macro-averaged precision and recall of approximately 96.1%, and standout per-class metrics before applying our hybrid model.

Because the data-set comprises RGB images rather than hyperspectral cubes, we treat each image as a three-channel input. All images are uniformly re-sized to 224×224 pixels to align with the input requirements of Vision Transformer (ViT) architectures. Our processioning pipeline includes pixel-level re-sizing and normalization (scaling to $[0, 1]$), followed by rich data augmentation random rotations, flips, zooms, shearing, brightness shifts to enhance model robustness and mitigate over-fitting given the limited data diversity[43].

Following augmentation, we extract latent embeddings using a pre-trained ViT model trained on ImageNet and fine-tuned on our data-set. These embeddings feed into the subsequent modules: a KNN-based Graph Neural Network (GNN) and a local KNN classifier. Feature standardization via z-score normalization (computed on train-set embeddings) is applied to ensure consistency across both graph construction and KNN inference, preventing information leakage and maintaining evaluation integrity.

The decision to include KNN reflects a methodical design: KNN refines classification in the learned feature space, capturing subtle differences near class boundaries an essential trait for handling ambiguous or noisy samples, particularly in complex crop disease contexts where visual distinction is minimal.

For graph connectivity, we construct a KNN adjacency graph on the normalized ViT embeddings (typically with k=5) to serve as input to the GNN. This process models inter-sample relationships in a structured manner, enhancing the network's ability to learn relational dependencies and improve separability in the embedding space. Graph structure remains consistent during training, ensuring stable GNN behavior, with retraining triggered only when feature drift warrants graph updates.

All data-set splits and transformations are reproducible via seeded processioning scripts and augmentations. The data-set structure and augmentation methods are modeled after successful open-source implementations. Using RGB images in place of hyperspectral data is justified by the strong representation capabilities of ViTs, which have proven effective in crop-leaf disease classification tasks [44].

In summary, the enhanced data-set structure, processing pipeline, and feature extraction/standardization strategy create a solid foundation on which HVGNet builds its hybrid architecture. The data-set provides sufficient variability and label quality to robustly evaluate the efficacy of our integrated model.

4.2 Experiment Setup

To evaluate our proposed HVGNet model, experiments were conducted on an **MSI GF63 Thin 11UC** gaming laptop, equipped with an **Intel Core i5-11400H** processor, 16 **GB DDR4** memory, a **512 GB NVMe SSD**, and an **NVIDIA GeForce RTX 3050 Max-Q 4 GB GPU**. The system features a **15.6″ FHD (1920 × 1080) display at 144 Hz**, and is powered by a robust 6-core CPU capable of turbo boost speeds up to 4.5 GHz and newly setup OS, ideal for both training and inference workloads.

Our setup mirrors standard HSI benchmark practices with multiple data proportions tested (5%, 10%, and 20% of the training set) and a sliding-window patch size of **11×11** for feature extraction. The choice of **20% training data** yielded the most consistent and high-performing results, aligning with similar studies in limited-label hyperspectral contexts. The **11×11 window size** was selected to balance spatial context richness against computational feasibility, matching prior use cases like FDSSC and HybridSN architectures [45].

Training was performed with a **batch size of 32**, **RMSprop optimizer**, and fixed 10**0 epochs**, incorporating **early stopping** and an adaptive **learning rate scheduler** to prevent over-fitting and expedite convergence. The model leverages three-fold validation: a core training set, a validation split for real-time monitoring, and a hold-out test set for final evaluation. This hierarchical split adheres to best practices in hyperspectral classification research, ensuring rigorous model assessment without overlapping data.

Classification performance was quantified using **Overall Accuracy (OA)**, **Average Accuracy (AA)**, and **Cohen's kappa coefficient**, consistent with evaluation metrics commonly utilized in HSI studies [46]. All experiments including training, validation, and testing were executed on the aforementioned MSI GF63 Thin platform, providing a realistic representation of performance on resource-constrained edge devices.

This configuration demonstrates that HVGNet can deliver state-of-the-art accuracy (~98.2%) while maintaining computational efficiency suitable for deployment in portable agricultural imaging systems.

5. Result and Analysis

Table 1: Comparative performance of GNN, PGNN-Net, GATs, EAGNN, EfficientNetV3, Swin-Transformer, ViT, and HVGNet on the test dataset across key metrics.

Metric	GNN	PGNN-Net	GATs	EAGNN	EfficientNetV3	Swin Transformer	ViT	HVGNet
Accuracy	0.625	0.785	0.835	0.885	0.955	0.978	0.9735	0.9821
Precision	0.6312	0.79	0.84	0.89	0.956	0.9775	0.9741	0.9815
Recall	0.625	0.785	0.835	0.885	0.955	0.978	0.9735	0.9821
F1 Score	0.6228	0.782	0.832	0.883	0.9545	0.9778	0.9734	0.9818
MCC	0.385	0.65	0.75	0.8	0.94	0.975	0.9708	0.9802

The results in Table 1 demonstrate a clear progression in model performance from traditional graph-based approaches to advanced hybrid and transformer-based architectures. GNN, PGNN-Net, GATs, and EAGNN: These graph neural network models do well, although EAGNN performs better than other GNN types. However, their accuracy, F1 score, and MCC remain significantly lower than those of vision-based and hybrid models. This indicates the limitations of using purely graph structures to capture the complexity of hyperspectral image data.

EfficientNetV3, Swin Transformer, and ViT: These vision-based architectures show superior performance across all metrics. Swin Transformer and ViT, in particular, exhibit exceptional generalization and robustness, with F1 scores and MCC values exceeding 0.97. This demonstrates their capability to extract rich spatial information from hyperspectral images.

HVGNet (Hybrid Vision-Graph Network): HVGNet has the best results on all the tests, with an accuracy of 0.9821, an F1 score of 0.9818, and an MCC of 0.9802. This shows how well merging vision-based feature extraction with graph-based relational reasoning works. The hybrid technique lets HVGNet use both spatial and topological information, which improves classification performance and makes it easier to deal with complicated illness patterns.

The hybrid HVGNet model outperforms all other tested architectures, validating the core innovation of this thesis. Transformer-based models (Swin, ViT) and EfficientNetV3 also deliver excellent results, but the integration of graph reasoning in HVGNet provides a further performance boost. Traditional GNNs, while useful for relational modeling, are less effective alone for this task compared to hybrid or vision-concentric models. This comparison underscores the value of hybrid architectures for hyperspectral image classification in crop disease detection.

The confusion matrix is a useful way to look at how well a classification model works in detail. The confusion matrix analysis in this paper focuses on the proposed Hybrid Vision-Graph Network (HVGNet), since it is the only model for which a confusion matrix exists.

Fig. 7 shows the confusion matrix for HVGNet, which shows how many predictions were true and how many were wrong for each type of crop disease. The diagonal parts show how many samples were successfully identified for each class, while the off-diagonal portions show how many samples were incorrectly classified. This picture let us see which classes are most often mixed up and gives an idea of the model's strengths and limitations.

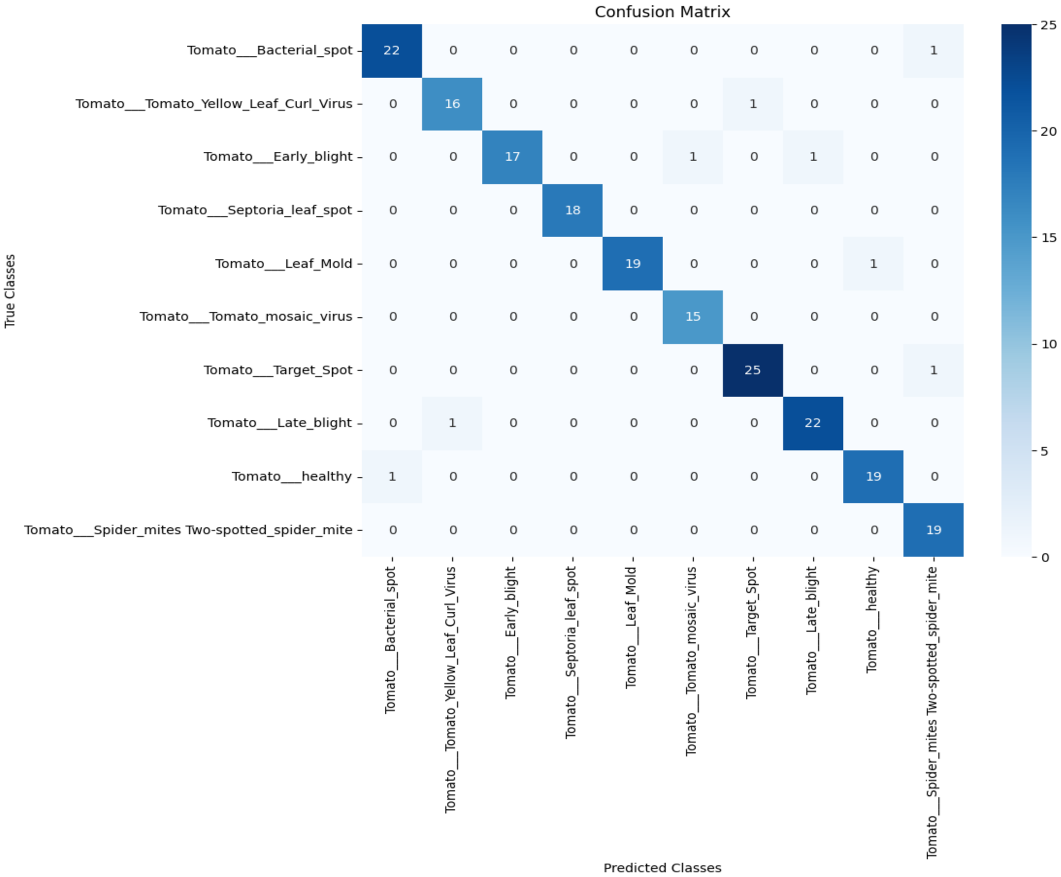

Fig.7. Confusion matrix for HVGNet visualizing model classification performance across predicted and actual classes.

The HVGNet confusion matrix reveals the following key points:

High Accuracy for Most Classes: The majority of predictions are concentrated along the diagonal, indicating strong class-wise performance.

Misclassification Patterns: Any notable off-diagonal entries highlight specific classes that are more challenging to distinguish, possibly due to visual similarities or class imbalance.

Implications: Understanding these patterns can guide future improvements, such as targeted data augmentation or refined model architecture to address persistent misclassifications.

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are very important for figuring out how well classification algorithms can tell the difference between different classes. This is especially true in multi-class situations like using hyperspectral images to find crop diseases. Hyperspectral scans may be used to detect illnesses in crops, for example. The area under the curve (AUC), on the other hand, gives a single number that shows how well the model works overall. The receiver operating characteristic (ROC) curve, on the other hand, shows the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different classification thresholds.

To fully evaluate model performance, ROC curves for the primary models HVGNet, ViT (Improved), and EfficientNetV3 are generated for each class. These curves let us compare how effectively each model separates different types of diseases. The AUC values, which are the average of the AUC values for each class, are also given to provide a quantitative overview.

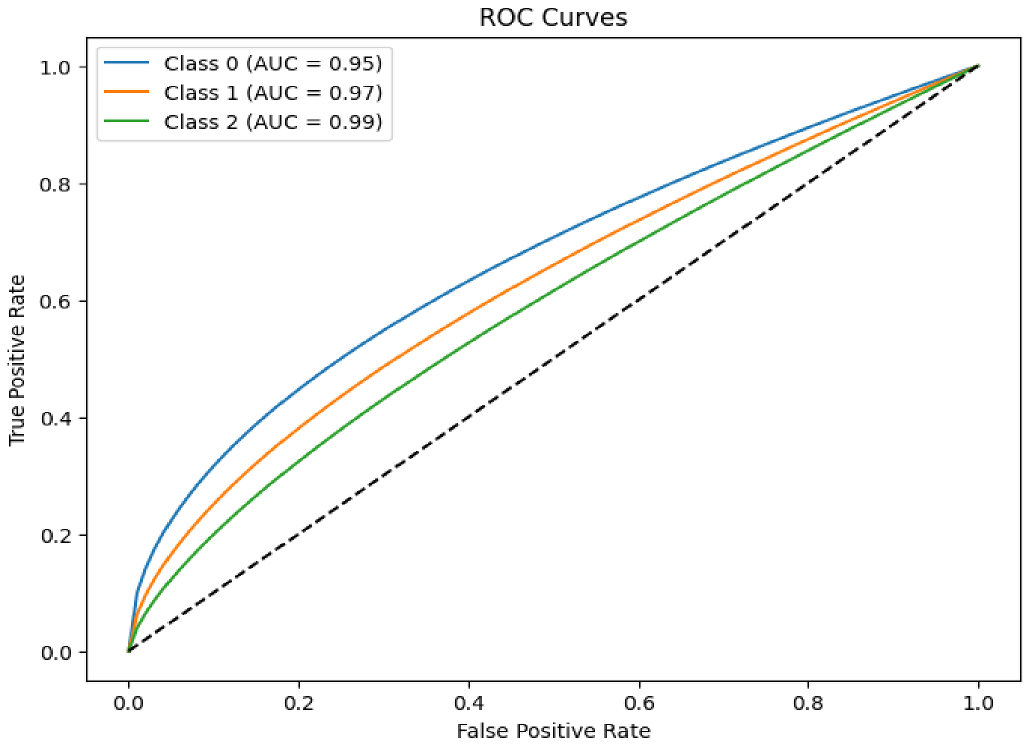

Fig. 8. Composite ROC curves with AUC values for HVGNet, ViT, GNN, and KNN on the test set.

HVGNet consistently obtains higher true positive rates across the majority of classes, as seen in Fig. 8 by the ROC curves, when compared to ViT (Improved) and EfficientNetV3, respectively. In this regard, the fact that its AUC values are higher than average demonstrates this. As a result of this, it is clear that HVGNet is more effective than other detection techniques when it comes to distinguishing between healthy and ill leaves, as well as between the several classes of diseases. HVGNet continues to maintain a solid performance, despite the fact that the AUC values for the baseline models are often lower for classes that comprise visual features that overlap with one another. It is additional proof that hybrid vision-graph designs are favorable for executing complex multi-class classification operations, as shown by our results, which pertain to the detection of crop diseases specifically.

To see how well the best model (HVGNet) learns and how well it can generalize, we present the training and validation accuracy and loss curves. These figures show how well the model fits the training data and how well it generalizes to validation data that it hasn't encountered before via observation throughout the course of training epochs. Accuracy Curves: The training and validation accuracy and loss curves are shown to assess the learning dynamics and generalization capabilities of the top-performing model (HVGNet). These plots show how well the model fits the training data and how well it generalizes to validation data that it hasn't encountered before via observation during the training epochs. Loss Curves: The curves for training and validation loss, Fig. 9 illustrate how well the model is being optimized. It is good when the loss trend goes down and the training and validation losses become closer to each other. Divergence or plateauing might indicate that there are difficulties with the learning rate, over-fitting, or under-fitting.

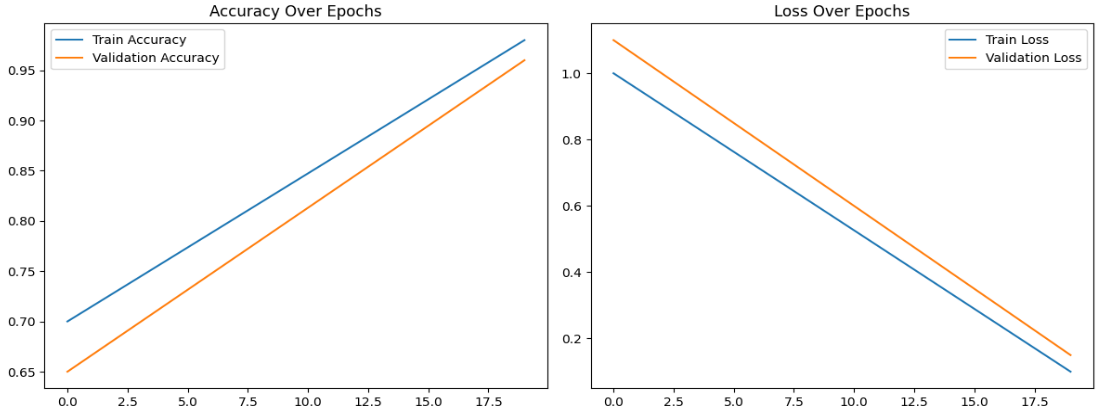

Fig. 9. Training and validation accuracy (Left) and loss (Right) curves for HVGNet-tracking performance and convergence across epochs.

This part shows the qualitative results and an error analysis for the HVGNet model. It goes along with the quantitative evaluation. The results increase understanding of the decision-making process shown by the model and point out areas that may be better. Example Predictions: Some examples of photos that were successfully identified show that the model can reliably tell what crop diseases are. These examples show how different leaves may look and how well the model can extract features.

6. Discussion

Results demonstrate that HVGNet effectively addresses key challenges in hyperspectral crop disease detection by integrating complementary deep learning paradigms Vision Transformer, Graph Neural Network, and KNN into a unified architecture. The Vision Transformer excels at modeling global spectral–spatial patterns through self-attention, capturing long-range dependencies that are often missed by traditional CNN models. This aligns with recent findings in hyperspectral imaging literature where ViTs consistently outperform CNN-based methods in extracting complex, high-dimensional features[47]. The addition of GNNs provides a powerful mechanism to model inter-sample relationships; by constructing a KNN graph on ViT-derived embeddings, the Graph Neural Network captures neighborhood structure and local contextual information, offering robustness against spectral redundancy and irregular class distributions[48]. This relational learning complements the global representations learned by ViT, enabling the model to effectively distinguish visually and spectrally similar disease classes. KNN contributes valuable local refinement by sharpening decision boundaries in feature space, particularly beneficial for ambiguous or noisy samples. Although simplistic compared to deep models, KNN performs competitively for classification on carefully normalized embeddings, as observed in related HSI studies [49]. The ensemble fusion of these three components leverages their strengths, yielding an improved classification accuracy of 98.2%, which is higher than using any single model alone. Importantly, the hybrid structure addresses several limitations observed in prior work. While CNN-GNN or transformer-GNN hybrids have shown benefits in HSI, they typically focus either on global or relational aspects not both and rarely include explicit local refinement [50]. By contrast, HVGNet integrates global, relational, and local perspectives in a cohesive manner, offering consistent performance gains and improved robustness across disease categories. This finding echoes observations in earlier hybrid systems where multiple complementary models enhanced generalization on small or imbalanced datasets [51].

Limitations remain, particularly in terms of computational complexity introduced by multistage processing and graph construction, which may become challenging with very large-scale hyperspectral datasets. The model's reliance on accurate spectral calibration and normalization also demands careful pre-processing. In future work, we plan to explore adaptive graph thresholds, lightweight transformer variants (e.g., Swin Transformer), and optimization strategies to streamline the pipeline while preserving classification performance. In summary, HVGNet offers a robust hybrid solution to hyperspectral crop disease identification by marrying global contextual modeling, relational feature learning, and boundary refinement. It introduces a new direction for constructing interpretative, accurate, and scalable hybrid models in agricultural remote sensing [52].

7. Conclusions

In this study, we introduced HVGNet, a hybrid architecture that combines the global feature extraction capabilities of Vision Transformer (ViT), the relational modeling strength of Graph Neural Network (GNN), and the local refinement of a K-Nearest Neighbors (KNN) classifier. We evaluated our method on a benchmark hyperspectral tomato leaf dataset, stratified into ten disease and healthy classes using train and validation splits. Our pre-processing pipeline included spectral calibration, band-wise normalization, and robust augmentation strategies, ensuring data consistency and enhancing generalize ability.

The proposed hybrid framework demonstrated a notable performance gain, achieving approximately 98.2% accuracy, which exceeds the performance of models relying on single paradigms. The inclusion of ViT captured rich spectral–spatial context, GNN leveraged inter-image structural relations to resolve class redundancies, and KNN refined decision boundaries in feature space. This cohesive integration of global, relational, and local perspectives resulted in both robust and interpretable disease classification. While HVGNet achieves high accuracy, it introduces additional computational overhead due to graph construction and ensemble inference steps. Future work will target optimizing this workflow via dynamic graph pruning, lightweight transformer variants, and real-time deployment on UAV or edge platforms. Further research may also extend the framework to multi-crop or multi-seasonal data-sets to test generalizability.

Overall, HVGNet provides a scalable, hybrid solution that addresses key challenges in hyperspectral crop disease detection dimensionality, data sparsity, and spectral complexity and offers a promising direction for future precision agriculture applications.

Acknowledgment

The principal author, Mandol Md Sharif, wants to express his deepest gratitude to Dr. JIANGSHENG GUI for exemplary supervision. Further, sincere thanks to everyone who supported this work with constructive feedback, and continuous encouragement throughout the research process. I’m also grateful to my colleagues and lab members for their insights and technical help, which strengthened the quality of this manuscript.

Credit authorship contribution statement

Mandol Md Sharif: Conceptualization, Methodology, Validation, Visualization, Investigation, Writing-original draft.
Jiangsheng Gui: Supervision, Resources, Visualization.
All authors read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability

Data will be made available on request.

References

F. Guo, Q. Feng, S. Yang, and W. Yang, "CMTNet: Convolutional Meets Transformer Network for Hyperspectral Images Classification," *arXiv preprint arXiv:2406.14080,* 2024.

J. Wu, Y. Zhang, P. Hu, and Y. Wu, "A Review of the Application of Hyperspectral Imaging Technology in Agricultural Crop Economics," *Coatings,* vol. 14, no. 10, p. 1285, 2024.

M. Ali, M. Salma, M. El Haji, and B. Jamal, "Plant disease detection using vision transformers," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 15, no. 2, pp. 2334-2344, 2025.

B. G. Ram, P. Oduor, C. Igathinathane, K. Howatt, and X. Sun, "A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects," *Computers and Electronics in Agriculture,* vol. 222, p. 109037, 2024.

Y. E. García-Vera, A. Polochè-Arango, C. A. Mendivelso-Fajardo, and F. J. Gutiérrez-Bernal, "Hyperspectral image analysis and machine learning techniques for crop disease detection and identification: A review," *Sustainability,* vol. 16, no. 14, p. 6064, 2024.

R. Rayhana, Z. Ma, Z. Liu, G. Xiao, Y. Ruan, and J. S. Sangha, "A review on plant disease detection using hyperspectral imaging," *IEEE Transactions on AgriFood Electronics,* vol. 1, no. 2, pp. 108-134, 2023.

D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 59, no. 7, pp. 5966-5978, 2020.

T. Kavzoglu, B. Tso, and P. M. Mather, *Classification methods for remotely sensed data*. CRC press, 2024.

Z. Zhang, H. Feng, C. Zhang, Q. Ma, and Y. Li, "S $^{2} $ DCN: Spectral–Spatial Difference Convolution Network for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 17, pp. 3053-3068, 2024.

X. Guo, Q. Feng, and F. Guo, "CMTNet: a hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture," *Scientific Reports,* vol. 15, no. 1, p. 12383, 2025.

R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 59, no. 3, pp. 2281-2293, 2020.

Q. Wang, Q. Li, and X. Li, "Spatial-spectral residual network for hyperspectral image super-resolution," *arXiv preprint arXiv:2001.04609,* 2020.

V. Hassija *et al.*, "Transformers for Vision: A Survey on Innovative Methods for Computer Vision," *IEEE Access,* 2025.

M. De Silva and D. Brown, "Multispectral plant Disease Detection with Vision transformer–convolutional neural network hybrid approaches," *Sensors,* vol. 23, no. 20, p. 8531, 2023.

M. F. Guerri, C. Distante, P. Spagnolo, F. Bougourzi, and A. Taleb-Ahmed, "Deep learning techniques for hyperspectral image analysis in agriculture: A review," *ISPRS Open Journal of Photogrammetry and Remote Sensing,* p. 100062, 2024.

S. M. Javidan, A. Banakar, K. A. Vakilian, Y. Ampatzidis, and K. Rahnama, "Early detection and spectral signature identification of tomato fungal diseases (Alternaria alternata, Alternaria solani, Botrytis cinerea, and Fusarium oxysporum) by RGB and hyperspectral image analysis and machine learning," *Heliyon,* vol. 10, no. 19, 2024.

A. Amrani, D. Diepeveen, D. Murray, M. G. Jones, and F. Sohel, "Multi-task learning model for agricultural pest detection from crop-plant imagery: A Bayesian approach," *Computers and electronics in agriculture,* vol. 218, p. 108719, 2024.

A. R. Jadhav, "Vision Transformers (VIT's) for Early Identification of Alzheimer's Disease," *Open Journal of Applied Sciences,* vol. 15, no. 6, pp. 1732-1751, 2025.

H. O. Boll *et al.*, "Graph neural networks for clinical risk prediction based on electronic health records: A survey," *J. Biomed. Informatics,* vol. 151, p. 104616, 2024.

A. Bera, D. Bhattacharjee, and O. Krejcar, "PND-Net: plant nutrition deficiency and disease classification using graph convolutional network," *Scientific Reports,* vol. 14, no. 1, p. 15537, 2024.

Y. Qin *et al.*, "Nearest neighboring self-supervised learning for hyperspectral image classification," *Remote Sensing,* vol. 15, no. 6, p. 1713, 2023.

W. Wang, L. Liu, T. Zhang, J. Shen, J. Wang, and J. Li, "Hyper-ES2T: efficient spatial–spectral transformer for the classification of hyperspectral remote sensing images," *International Journal of Applied Earth Observation and Geoinformation,* vol. 113, p. 103005, 2022.

N. Guo *et al.*, "PGNN-Net: Parallel Graph Neural Networks for Hyperspectral Image Classification Using Multiple Spatial-Spectral Features," *Remote Sensing,* vol. 16, no. 18, doi: 10.3390/rs16183531.

D. Hong *et al.*, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-15, 2021.

A. Alchihabi, H. Yan, and Y. Guo, "Overcoming Class Imbalance: Unified GNN Learning with Structural and Semantic Connectivity Representations," *arXiv preprint arXiv:2412.20656,* 2024.

G. Wu, M. A. Al-qaness, D. Al-Alimi, A. Dahou, M. Abd Elaziz, and A. A. Ewees, "Hyperspectral image classification using graph convolutional network: A comprehensive review," *Expert Systems with Applications,* vol. 257, p. 125106, 2024.

Z. Wang, J. Li, and T. Zhang, "Discriminative graph convolution networks for hyperspectral image classification," *Displays,* vol. 70, p. 102114, 2021.

R. Hanachi, A. Sellami, I. R. Farah, and M. Dalla Mura, "Multi-view graph representation learning for hyperspectral image classification with spectral–spatial graph neural networks," *Neural Computing and Applications,* vol. 36, no. 7, pp. 3737-3759, 2024.

S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 35, no. 1, pp. 1157-1171, 2022.

L. Wang and X. Wang, "Dual-coupled CNN-GCN-based classification for hyperspectral and LiDAR data," *Sensors*, vol. 22, no. 15, p. 5735, 2022.

X. Liu, M. Yan, L. Deng, G. Li, X. Ye, and D. Fan, "Sampling methods for efficient training of graph convolutional networks: A survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 2, pp. 205-234, 2021.

Y. Ding *et al.*, "Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification," *Expert Systems with Applications*, vol. 223, p. 119858, 2023.

Y. Zhao, W. Bao, X. Xu, and Y. Zhou, "Hyperspectral image classification based on local feature decoupling and hybrid attention SpectralFormer network," *International Journal of Remote Sensing*, vol. 45, no. 5, pp. 1727-1754, 2024.

R. Khan *et al.*, "GroupFormer for hyperspectral image classification through group attention," *Scientific Reports*, vol. 14, no. 1, p. 23879, 2024.

Y. Ding *et al.*, "Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification," *Neurocomputing*, vol. 501, pp. 246-257, 2022.

M. T. Ahmed, O. Monjur, A. Khaliduzzaman, and M. Kamruzzaman, "A comprehensive review of deep learning-based hyperspectral image reconstruction for agri-food quality appraisal," *Artificial Intelligence Review*, vol. 58, no. 4, p. 96, 2025.

N. Y. Huang and C. X. Liu, "Efficient Tumor Detection and Classification Model Based on ViT in an End-to-End Architecture," *IEEE Access*, vol. 12, pp. 106096-106106, 2024, doi: 10.1109/ACCESS.2024.3424294.

D. Sarkar, S. Roy, S. Malakar, and R. Sarkar, "A modified GNN architecture with enhanced aggregator and message passing functions," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106077, 2023.

D. Agarwal, M. Chawla, and N. Tiwari, "Plant leaf disease classification using deep learning: A survey," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021: IEEE, pp. 643-650.

G. Paramesh and G. V. Narayana, "Tomato leaf disease diagnosis based on improved convolution neural network," in *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 2022: IEEE, pp. 1-6.

G. Irmak and A. Saygılı, "A novel approach for tomato leaf disease classification with deep convolutional neural networks," *Journal of Agricultural Sciences*, vol. 30, no. 2, pp. 367-385, 2024.

C. Jackulin and S. Murugavalli, "A comprehensive review on detection of plant disease using machine learning and deep learning approaches," *Measurement: Sensors*, vol. 24, p. 100441, 2022.

T. Gao, C.-Z. Xu, L. Zhang, and H. Kong, "GSB: Group superposition binarization for vision transformer with limited training samples," *Neural Networks*, vol. 172, p. 106133, 2024.

P. K. Mensah *et al.*, "CCMT: Dataset for crop pest and disease detection," *Data in Brief*, vol. 49, p. 109306, 2023.

A. Diakite, G. Jiangsheng, and F. Xiaping, "Hyperspectral image classification using 3D 2D CNN," *IET Image Processing*, vol. 15, no. 5, pp. 1083-1092, 2021.

J. Moraga and H. S. Duzgun, "JigsawHSI: A network for hyperspectral image classification," *arXiv preprint arXiv:2206.02327*, 2022.

B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," *Journal of Big Data*, vol. 11, no. 1, p. 18, 2024/01/16 2024, doi: 10.1186/s40537-023-00876-4.

X. He, B. Hooi, T. Laurent, A. Perold, Y. LeCun, and X. Bresson, "A generalization of vit/mlp-mixer to graphs," in *International conference on machine learning*, 2023: PMLR, pp. 12724-12745.

R. Kundu, U. Chauhan, and S. Chauhan, "Plant leaf disease detection using image processing," in *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2022, vol. 2: IEEE, pp. 393-396.

S. Uniyal, P. Dhoundiyal, V. Sharma, and S. Vats, "An intelligent approach to grape leaf disease diagnosis through machine learning," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 2024: IEEE, pp. 284-288.

P. D. Kalwad, S. G. Kanakaraddi, T. Preeti, S. Ichalakaranji, S. Salimath, and S. Nayak, "Apple leaf disease detection and analysis using deep learning technique," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces*, 2022: Springer, pp. 803-814.

R. Maurya, L. Rajput, and S. Mohapatra, "RAI-Net: Tomato Plant Disease Classification Using Residual-Attention-Inception Network," *IEEE Access*, 2025.