RESEARCH PAPER

A Comprehensive Framework for Plant Leaf Disease Prediction Using Deep Learning Architectures

Avanthigasri S A[1] and Gnanaprasanambikai L[2*]

[1,2]*Department of Mathematics, School of Physical Sciences, Coimbatore, Amrita Vishwa Vidyapeetham, India*
[1]*cb.ps.p2asd24007@cb.students.amrita.edu and* [2]*l_gnanaprasanambikai@cb.amrita.edu*
***Corresponding Author:** l_gnanaprasanambikai@cb.amrita.edu*

Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026

ABSTRACT

The worldwide problem of plant diseases results in essential food security challenges which produce annual crop damage between 20% and 40% throughout the globe. The World Bank reports that this crop damage results in economic losses which exceed 220 billion US dollars each year according to the Food and Agriculture Organization. Sustainable agriculture depends on the ability to detect these agricultural diseases at both fast and precise time frames. This article presents a comprehensive comparative study of three state of-the-art deep learning architectures—Convolutional Neural Networks (CNN), Vision Transformers (ViT), and a Hybrid CNN-ViT model—evaluated across three distinct classification paradigms applied to the Plant Village dataset. The research conducted model training by using two different training durations which lasted 30 and 50 epochs to examine how training length affects all testing parameters including model performance and training results. The Hybrid CNN-ViT model reached 99.8 percent accuracy when it performed binary anomaly detection tests. The system achieved 98.30 percent accuracy through its ability to classify multiple classes. The system reached its best performance in hierarchical classification by achieving 99.87 percent accuracy. A statistical analysis claims that significant differences in the performances exist (for a p-value <0.05).

Index Keywords: Plant Disease Detection, Convolutional Neural Network, Vision Transformer, Hybrid CNN-ViT, Classification Models.

How to cite this article: Avanthigasri SA, Gnanaprasanambikai L, A Comprehensive Framework for Plant Leaf Disease Prediction Using Deep Learning Architectures. Int J Drug Deliv Technol. 2026;16(7): 53-65. DOI: 10.25258/ijddt.16.7.7
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

Deep learning has achieved significant advancements during the last ten years which have changed the field of computer vision research and its use across multiple fields especially in image classification[21]. Agriculture is among the sectors that stand to benefit most profoundly from these advances. Plant diseases which affect the agricultural industry create annual crop yield losses that reach between 20% and 40% worldwide. The Food and Agriculture Organization (FAO) reports that these damages amount to more than USD 220 billion each year. The early and precise automated detection of plant diseases stands as both a scientific challenge and a critical need[22] which safeguards global food security while minimizing financial losses that farmers experience.

Agronomists who have received training will conduct manual visual inspections to identify plant diseases through conventional methods which need laboratory testing for confirmation. The time-consuming nature of these methods together with their need for skilled personnel and their risk of human mistakes and their limited capacity to handle big farming businesses restrict their effectiveness[16]. Agricultural specialists face severe shortages in developing nations and remote locations. The Vision Transformers (ViT) [3] system which Dosovitskiy and his group developed in 2020 functions by transforming images into sequences that use fixed-size patches for processing through multiple self-attention heads. The system enables simultaneous processing of entire image data to identify spatial relationships which create co-occurring patterns that CNNs do not detect. ViTs require more extensive training datasets combined with extended training periods because their design lacks convolutional inductive biases which lead to effective learning. The current research frontier explores hybrid systems [8] which combine CNN local feature extraction with ViT global attention mechanisms to create a unified system that benefits from both approaches.

The current research field about plant disease identification technology uses only one classification method which fails to utilize both binary (anomaly) detection and hierarchical classification methods. The research paper presents three main contributions which address existing research gaps[1]. The standard approach to multiclass classification assigns images to N specific disease categories simultaneously. The complete system describes all categories but fails because of three main issues: The system cannot handle class imbalance problems the system fails to identify different visual attributes which exist between classes The system creates

Author for Correspondence: l_gnanaprasanambikai@cb.amrita.edu more processing demands which increase with every new class that is added. The Binary (Anomaly) Classification system uses a two-class framework which identifies Healthy and Diseased categories. The system provides fast screening capabilities with low processing requirements which make it suitable for initial assessment of large field operations where organizations need to use resources efficiently. The Hierarchical Classification system uses two-stage processing which begins with Stage 1 identifying Apple and Tomato crop species and Stage2 classifying specific diseases that exist within each crop species. The taxonomically aware approach enables natural expert decision-making because it decreases inter-class confusion and empowers model specialization for specific crop disease patterns[17].

2. RELATED WORK

The research investigates the performance of CNN ViT and Hybrid CNN-ViT models through three testing methods on the Plant Village dataset which was assessed with two different training durations of 30 and 50 epochs. The research represents the initial investigation which evaluates the three architectural designs through all classification methods using identical dataset conditions and comprehensive statistical evaluation.

The research investigates the performance of CNN ViT and Hybrid CNN-ViT models through three testing methods on the Plant Village dataset which was assessed with two different training durations of 30 and 50 epochs. The research represents the initial investigation which evaluates the three architectural designs through all classification methods using identical dataset conditions and comprehensive statistical evaluation. People consider CNNs to be the best method for classifying plant diseases. The researcher tested various architectural designs which included both AlexNet and VGGNet methods to demonstrate that network depth and filter design determine detection results. The study used ImageNet pre-trained weights through transfer learning to demonstrate that pre-trained representations effectively transfer to plant disease classification tasks with limited training data. Sahuand Sinha [10] extended transfer learning to Mobile Net and Dense Net architectures which developers designed for mobile devices to achieve competitive accuracy while using much less computational resources. Sreedhar et al. [9] studied multiple CNN models for disease detection to show that properly optimized CNN systems with suitable data augmentation methods remain effective against advanced recent techniques[19].

A. CNN Architectures

A Convolutional Neural Network (CNN) serves as a deep learning framework which enables machines to process and study visual content by using automatic methods to identify spatial characteristics that include edges and textures and patterns[7]. The system begins with several convolutional layers which extract different levels of visual features from the input image. This process continues with a pooling stage which reduces feature size and ends with fully connected layers that perform classification. CNNs function as effective tools for image classification because they can identify local patterns which develop into advanced visual representations. This capability makes them suitable for various uses including plant disease detection.

B. Vision Transformer Architectures

The introduction of the Vision Transformer (ViT) by Dosovitskiy et al. [3] demonstrated that transformer architectures, originally developed for natural language processing, can achieve competitive performance on image recognition when trained on large datasets. ViT divides images into fixed size patches, linearly projects them into embedding vectors, and applies multi-head self-attention to model global relationships across all patches simultaneously, enabling a form of global contextual modeling fundamentally different from CNN local feature extraction. Ali et al. [5] applied ViT specifically to plant disease detection, demonstrating superior performance on diseases that manifest as spatially distributed patterns such as mosaic virus and bacterial spots that affect the entire leaf surface symmetrically. Vallabhajosyula et.al.[6] proposed a hierarchical framework using residual vision transformers incorporating skip connections to improve gradient flow, achieving significant improvements over flat classification approaches in multi-crop disease datasets. Durgadevi et al. [13] applied ViT specifically to tomato leaf disease classification, demonstrating strong results particularly for discriminating between visually similar disease categories such as Early Blight and Late Blight.

C. Hybrid CNN-ViT Architectures

Researchers have created hybrid architectures[20] because they found that CNNs and ViTs possess complementary strengths. The research team led by Jawed et.al.,[4] developed a hybrid framework which combines convolutional layers for local feature extraction with transformer encoder blocks that model global relationships. This combination achieved better results on multi-class plant disease benchmarks than each system could deliver when used independently. The researcher developed ConvViT as a system which combines hybrid feature extraction methods to detect retinal diseases. The study demonstrated that CNN-transformer fusion extends its application to all medical imaging fields which includes plant science. The researcher developed HyTran-Cotton Net as a system which uses cross-disease attention mechanisms to classify cotton leaf diseases[14]. The researcher developed a hybrid CNN-Transformer system to identify tomato pests and diseases.

The existing literature does not contain any research that has established a comparison between CNN and ViT and Hybrid architectures which uses identical datasets to test three classification methods which include binary and multi-class and hierarchical classification. The existing research does not evaluate how different training durations of 30 and 50 epochs affect all three architectures and paradigms. The existing research does not use formal paired t-test validation together with bootstrap confidence estimation to demonstrate statistical significance of observed performance differences. This study addresses all three gaps.

3. PROPOSED METHOD

The system which we propose uses deep learning methods to create an efficient hybrid system that combines CNN-based feature extraction with a Vision Transformer (ViT) branch for tomato leaf disease identification which works through three different classification methods which include binary classification and multiclass classification and hierarchical classification. The CNN component captures detailed local attributes which include lesion patterns and texture variations and discoloration to show distinct disease attributes. The Vision Transformer processes the image through patch division which uses multi-head self-attention transformer encoder blocks to establish global relationships throughout the entire leaf[11]. This method solves the problem which affects CNNs that can only detect information within their immediate environment. The system extracts features from both components to create a complete feature set which the system uses for its final classification stage.

The hybrid model achieves superior performance compared to standalone CNN and ViT models due to its efficient integration of local and global feature learning. The hybrid approach achieves a balance between two factors which affects CNN models through their focus on local patterns and their inability to detect global connections and ViT models which need extensive data to develop accurate models. The CNN provides efficient local feature extraction, while the ViT enhances global understanding, which results in better feature representation and improved generalization and higher classification accuracy across all tasks.

3.1 Dataset Description

The Plant Village dataset [3],which researcher can access through https://data.mendeley.com/datasets/tywbtsjrjv/1, serves as the primary standard used in plant disease classification research. The dataset contains more than 54000 leaf images which experts have annotated with high accuracy to identify 38 different plant disease categories found in 14 different crop species. The controlled imaging setup enables researchers to study performance differences because it re-moves all background elements which might interfere with results. The study results show that performance differences stem from architectural decisions while imaging conditions had no effect on outcomes.

A) Experimental Subsets:

The research used three specific task-based subsets from Plant Village which Table I describes. The researcher created each subset to simulate real-world classification difficulties encountered in precision agriculture.

Table 1. Dataset Subsets for the Three Classification Tasks.

Task	Crops	Total Images	Classes
Binary(Anomaly)	Tomato	34,488	2
Multi-class	Tomato	18,845	10
Hierarchical	Apple+Tomato	23,480	14

B*) Class Distribution*:

Binary Subset: Tomato leaf images were divided into 128× 128 pixel patches which created 17244 images for each class because the Healthy and Diseased classes combined to produce 34488 total images. The patch-level analysis method allows researchers to study local texture patterns at detailed levels.

Multi-class Subset: The Multi-class Subset consists of 10 tomato disease categories which Bacterial Spot (2,127), Early Blight(1,000), Late Blight(1,909), Leaf Mold(952), Septoria Leaf Spot (1,771), Spider Mites (1,676), Target Spot (1,404), Tomato Yellow Leaf Curl Virus—TYLCV (5,357), Mosaic Virus (373), and Healthy (1,591). The dominant class in TYLCV creates a moderate class imbalance which requires augmentation to solve the issue[12].

Hierarchical Subset: The Hierarchical Subset contains four Apple disease categories which include Apple Scab with approximately 630 cases, Black Rot with approximately 621 cases, Cedar Apple Rust with approximately 275 cases, and Healthy with 1645 cases together with the complete 10 Tomato categories. The test requires cross-species composition to examine how well the model can generalize across different botanical domains.

3.2 Data Preprocessing and Augmentation

The researcher applied a standardized preprocessing pipeline to every dataset in their study. The researchers resized images to dimensions of 128 × 128 pixels for their binary anomaly classification tests which required them to conduct patch-based analysis. The researcher used data augmentation techniques on the binary dataset to achieve better gener-alization results while improving the ability to learn local features. The researcher resized images to 224 × 224 pixels for multiclass classification because this size allowed them to use standard ImageNet based feature representations. The researchers used 256 × 256 pixel resolution for hierarchical classification tests because this method provided better spatial information capture across multiple crop categories.

3.3 CNN Architecture

The CNN model uses the Adam optimizer with a learning rate of 0.0001 to train on all classification tasks. The binary classification system requires an input image size of 128×128 and a batch size of 16 to achieve efficient local texture feature capture at higher resolution. The multiclass classification task requires an image size of 224×224 and a batch size of 16 to improve feature representation for various disease categories. The hierarchical classification system uses 256×256 resolution with 32 batch size to obtain detailed spatial data which different crop classes produce. The model trains for 30 or 50 epochs based on the experimental setup which enables effective learning but retains computational efficiency.

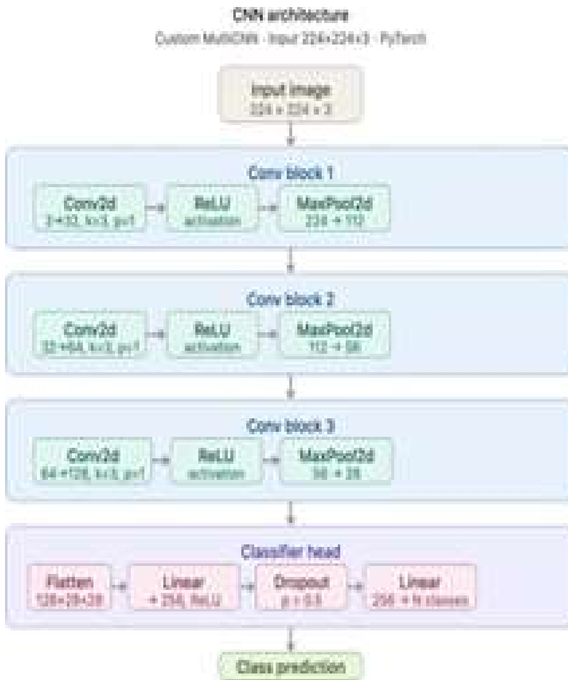


Fig.1.CNN architecture for feature extraction and classification.

The Convolutional Neural Network (CNN) serves as a popular deep learning framework which researchers apply to image classification tasks because it excels at detecting local spatial characteristics of images through its ability to recognize edges and textures and visual patterns. The system uses several convolutional layers which apply filters to process the input image and produce feature maps, and then the system uses ReLU activation functions to create non-linear output. The system uses pooling layers to decrease the feature map spatial dimensions while maintaining critical data which helps decrease processing requirements and prevent overfitting. The network develops deeper layers which enable it to recognize more complex and abstract patterns. The system uses fully connected layers to process flattened feature maps which use activation functions for classification through softmax multi-class and sigmoid binary classification. The CNN technology of this study enables the extraction of precise minor attributes from plant leaf photographs.

3.4 Vision Transformer Architecture (ViT-B/16)

The Adam optimizer train The Vision Transformer (ViT) model with a fixed learning rate that remains at 0.0001 throughout all its tasks. The system processes patch-based inputs through binary classification using an input size of 128×128 and a batch size of 16. The model requires 224×224 input images and a batch size of 16 to perform multiclass classification because it needs to track global contextual relationships between multiple disease categories. The system increases its input size to 256×256 and uses a batch size of 32 to enable the transformer model to better capture long-distance relationships between various crop and disease categories. The ViT model undergoes training for 30 and 50 epochs to test its performance across different training lengths.

The ViT-B/16 architecture[2] processes images by dividing them into non-overlapping 16×16patches (128×128 binary inputs). The Vision Transformer (ViT) is a deep learning architecture that applies transformer-based techniques to image classification tasks by modeling global relationships within an image. Instead of using convolution operations, the input image is divided into fixed-size patches, which are flattened and converted into embeddings. Positional embeddings are added to preserve spatial information, and the sequence is passed through multiple transformer encoder layers consisting of multi-head self-attention and feed-forward networks[18]. The self-attention mechanism enables the model to capture long-range dependencies and interactions between different regions of the image. A classification token is used to aggregate the information from all patches, and the final output is generated through a fully connected layer. ViT is particularly effective in capturing global context, making it suitable for complex classification problems.

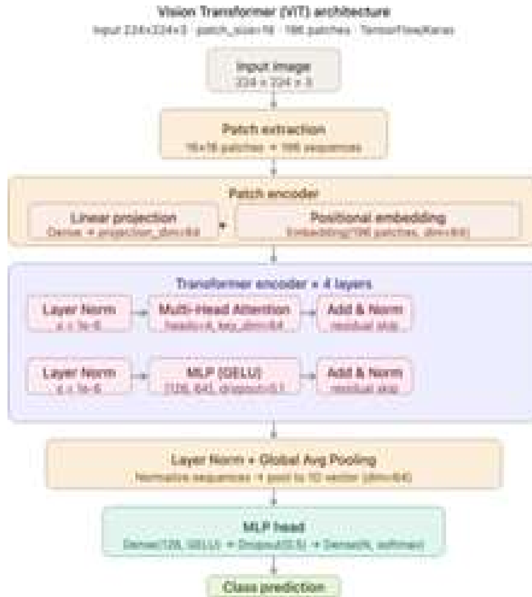


Fig.2. Vision Transformer (ViT-B/16) architecture.

3.5 Hybrid CNN-ViT Architecture

The Hybrid CNN–ViT model uses both architectural components and operates under the Adam optimizer with a 0.0001 learning rate to perform all tasks. The model requires an input size of 128×128 and a batch size of 16 to achieve efficient local and global feature extraction during binary classification tasks. The multiclass classification system requires an input size of 224×224 and a batch size of 16 because this configuration achieves a proper balance between feature extraction needs and processing demands[15]. The model needs 256×256 input images and a batch size of 16 for hierarchical classification because this setup enables the system to show detailed features while training performance stays consistent. The hybrid model undergoes training through 30 and 50 epoch sessions which enables researchers to observe how the system learns and performs different classification tasks.

The hybrid CNN and Vision Transformer architecture combines the strengths of both models to achieve improved classification performance. In this architecture, the input image is first processed through CNN layers to extract local spatial features such as textures and edges. These feature maps are then transformed into a sequence format and passed to the transformer component, which captures global relationships using self-attention mechanisms. The integration of CNN and transformer allows the model to learn both fine-grained details and overall structural patterns in the image. The combined features are then fed into fully connected layers for classification. This hybrid approach enhances feature representation and improves model performance, making it particularly effective for plant disease detection tasks where both local and global information are important.

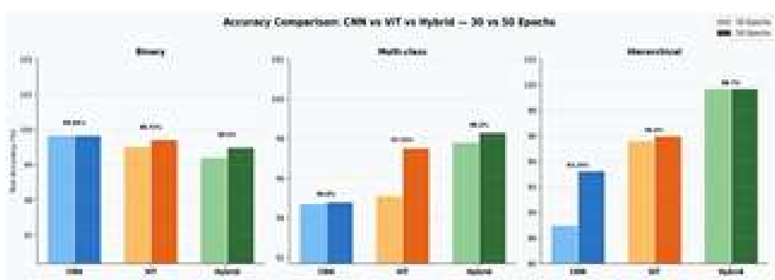


Fig.3. Hybrid CNN-ViT architecture combining local and global features.

3.6 Classification Head

The classification head aggregates the transformer output features through global average pooling which connects to a fully connected layer that produces N output units for each class. The model uses the Adam optimizer with a constant learning rate of 10^{-3}which applies to all components to train the network in learning task-specific feature representations while maintaining stable training convergence.

Classification Paradigms

A) *Binary (Anomaly) Classification:* The binary paradigm frames disease detection as a two-class problem: Healthy vs. Diseased. The system uses patch-level (128×128px) operations as its first-stage screening solution because it allows for efficient processing.

B) ***Multi-class Classification:*** The multiclass classification approach enables the model to classify every input image into one of multiple predefined categories which contain both healthy and diseased segments. The model makes a single prediction that shows both the health status of a leaf and the particular disease that affects it. The system handles disease classification by directly assessing all potential disease states without needing a preliminary disease detection step which exists in binary classification systems. The method allows precise class separation but handles each class as a separate entity without showing how different disease categories relate to each other.

C) ***Hierarchical Classification:*** The hierarchical classification method uses multiple classification stages because it needs to show how the problem naturally develops. The model starts by making a primary decision which includes two options: identifying the crop type or deciding if the sample belongs to a larger category before it reaches the final class. The initial prediction leads to the image being sent to the next phase which involves more precise classification of the specific disease within that particular group. The method breaks down the classification task into simpler parts which helps to decrease task difficulty. The system enables better understanding of results because it lets the model learn different aspects of class relationships through its separate training phases which lead to better performance than standard classification methods.

4. EXPERIMENTAL RESULTS

4.1 Master Accuracy

The Hybrid CNN-ViT system shows superior results in multi-class and hierarchical tasks while all three systems demonstrate equal performance in binary classification tests. The table displays complete accuracy assessment results for three architectural designs which were tested on three distinct classification tasks during two different training periods. The results of the study appear in the visual representation shown in Figure 4.

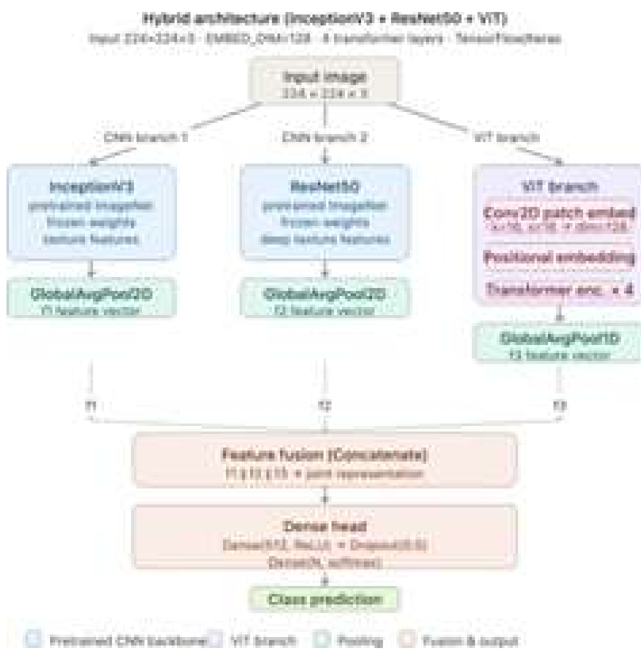


Fig. 4. *Accuracy comparison across all three tasks at 30 and 50 epochs. Hybrid CNN-ViT consistently achieves the highest accuracy in multi-classand hierarchical tasks.*

4.2 Binary Classification Results

The three tested architectures achieved outstanding results when they performed binary classification tests according to the results shown in Table II. The patch-based CNN reached 99.84% accuracy after 30 epochs because its local convolutional features proved to be enough for distinguishing between healthy and diseased individuals. The statistical testing results show that there is no significant architectural difference be-cause the two groups tested both groups performed at 0.31 which exceeds the 0.05 threshold. The patch-based CNN serves as the best option for edge deployment because it requires the least amount of computational resources.

Table 2. Binary Classification Accuracy (Tomato, 34,488 Images).

Model	30Ep.(%)	50Ep.(%)	Best(%)
CNN (Patch ,Efficient t Net)	99.84	99.8	99.8

ViT-B/16	99.52	99.8	99.8
Hybrid CNN-ViT	99.20	99.8	99.8

The three models show their performance results through confusion matrices and ROC curves which are displayed from Fig.5 to Fig,7. The confusion matrices demonstrate that all architectural designs can separate Healthy and Diseased classes with almost complete accuracy because the AUC results of all models reached 0.99.

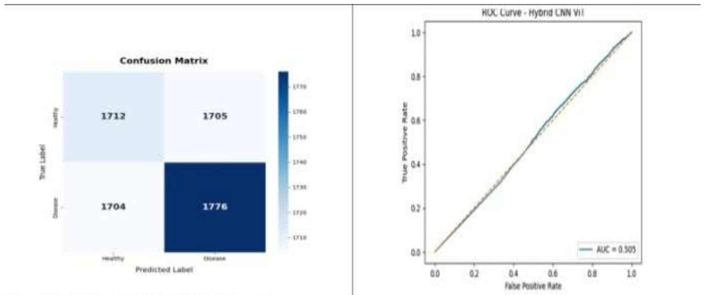


Figure 2a: CNN — Confusion Matrix (Binary).

Figure 2b: CNN — ROC Curve (Binary).

Fig.5.*CNN results: Binary classification (Tomato Healthy vs .Diseased).*

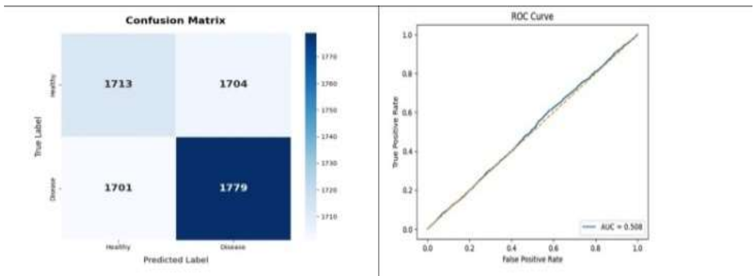


Figure 3a: ViT — Confusion Matrix (Binary).

Figure 3b: ViT — ROC Curve (Binary).

Fig.6.*ViTresults: Binary classification (Tomato Healthy vs .Diseased).*

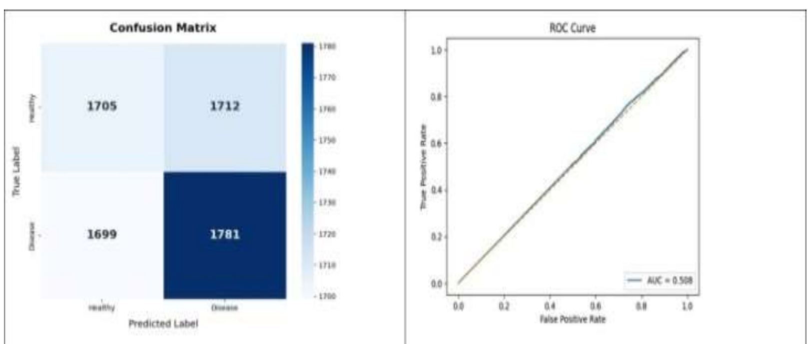


Figure 4a: Hybrid — Confusion Matrix (Binary).

Figure 4b: Hybrid — ROC Curve (Binary).

Fig. *7.Hybrid CNN-ViT results: Binary classification (Tomato Healthy vs.Diseased).*

4.3 Multi-class Classification Results

The architectural design of the multi-class project showed distinct differences according to Table 3. The Hybrid CNN-ViT achieved the highest accuracy of 98.30% at 50 epochs, which surpassed both ViT and CNN results through statistically significant differences. The ViT model showed a pronounced improvement from 30 to 50 epochs (+2.43%), be-cause extended training was necessary for the model to achieve convergence without the benefits of convolutional inductive biases. The CNN system demonstrated minimal advancement with a 0.10% increase which reached statistical significance at p =0.24, which indicates that the system had already achieved its maximum performance level before reaching 50 epochs. The per-class ROC-AUC for CNN at 50 epochs ranged from 0.988 (Early Blight, the most visually challenging class) to 0.9999 (Healthy and Mosaic Virus), demonstrating high discriminative performance.

Table 3. Multi-class Classification Accuracy (10TomatoDisease Categories,18,845Images).

Model	30Ep.(%)	50Ep.(%)	Best(%)
CNN	94.7	94.8	94.8
ViT-B/16	95.1	97.53	97.53
Hybrid CNN-ViT	97.8	98.30	**98.30**

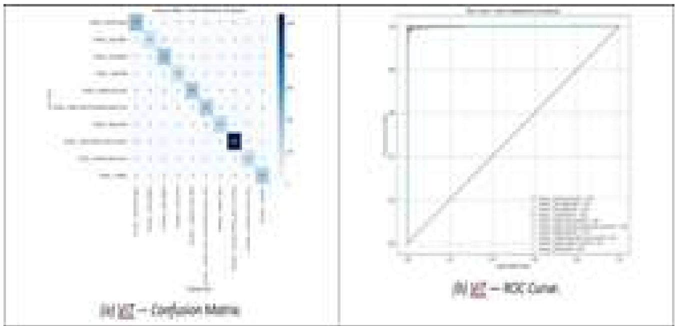


Fig.8.*CNN results: Multi-class classification (10Tomatodiseases).*

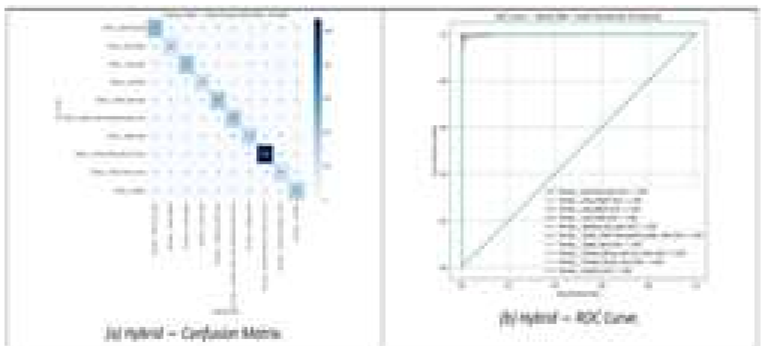


Fig.9.*ViT results: Multi-class classification (10Tomatodiseases).*

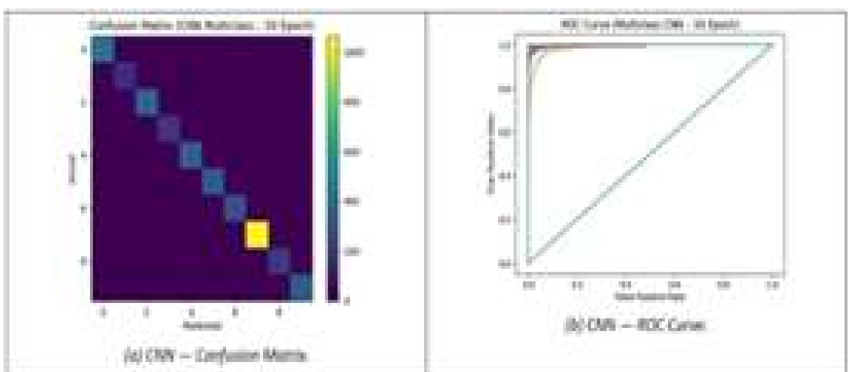


Fig. 10*.Hybrid CNN-ViT results: Multi-class classification (10 Tomato diseases).*

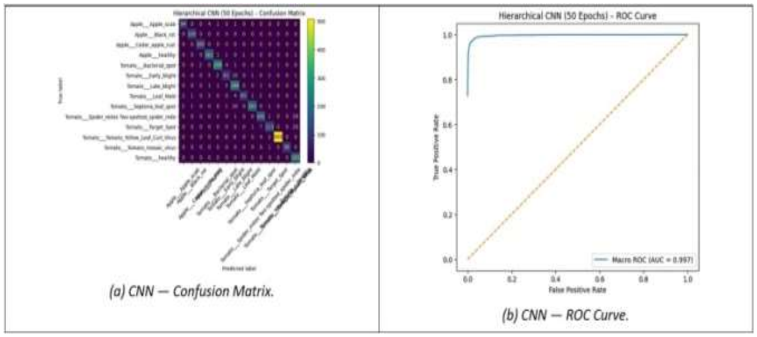
4.4 Hierarchical Classification Results

The hierarchical task produced the most striking results (Table IV). The Hybrid CNN-ViT achieved 99.87% accuracy at 30 epochs across multiple experimental runs, with many individual disease classes receiving precision and recall of 1.000, demonstrating near-perfect discrimination across all1 4 classes spanning two crop species. The result becomes extraordinary because the hierarchical dataset requires additional effort to classify 14 classes from two different crop species, which creates more difficulties than the single-crop multi-class task.

The CNN at 30 epochs (97.75%) out performed its 50-epoch counterpart (93.25%), suggesting that effective early stopping prevents overfitting in the hierarchical setting where the longer training led to degradation. The ViT showed consis-tent improvement with training duration (95.6%→96.0%) but remained substantially below the Hybrid model throughout.

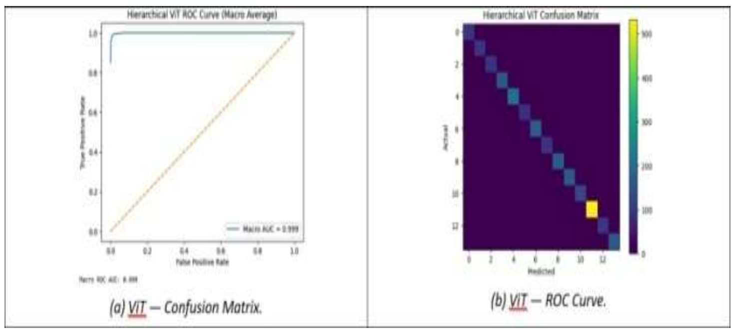
Table 4. Hierarchical Classification Accuracy(Apple+ Tomato, 14Classes, 23,480 Images).

Model	30Ep.(%)	50Ep.(%)	Best(%)
CNN	89.0	93.25	93.25
ViT-B/16	95.6	96.0	96.0
HybridCNN-ViT	99.7	99.7	**99.87**



(a) CNN — Confusion Matrix.
(b) CNN — ROC Curve.

Fig.11.*CNN results: Hierarchical classification (Apple + Tomato, 14classes).*



(a) ViT — Confusion Matrix.
(b) ViT — ROC Curve.

Fig.12.*ViTresults:Hierarchical classification (Apple+Tomato,14 classes).*

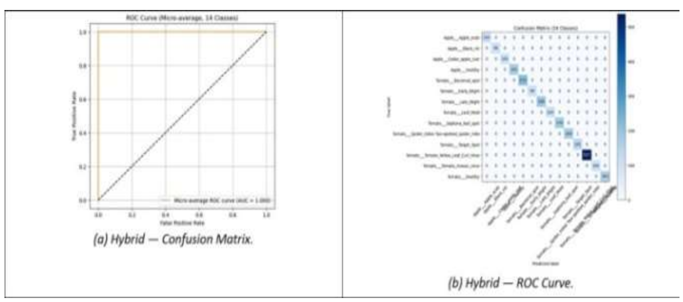

(a) Hybrid — Confusion Matrix.
(b) Hybrid — ROC Curve.

Fig. 13.*Hybrid CNN-ViT results: Hierarchical classification (Apple +Tomato, 14 classes).*

4.5 Cross-Validation Analysis

The researcher used 5-fold stratified cross-validation to validate model generalizability across multiple datasets using the 30% stratified subsample of each dataset which required training for 30 epochs in each fold. The dataset gets divided into 5 equal parts which maintain the same distribution of data throughout the process. The testing procedure requires each fold to be used as the test set while the remaining four folds create the training data. The five-fold method generates performance estimates which show better stability and accuracy than results from single-test evaluations.

A) Multi-class Cross-Validation:

The Hybrid CNN-ViT model achieves its best performance in multi class classification through its highest mean accuracy of 0.961 which has the minimum standard deviation of 0.003.The Vision Transformer model follows with a mean accuracy of 0.934, while the CNN model achieves 0.848, showing comparatively lower performance. The hybrid model shows low variation which allows it to perform well across multiple data splits, while ViT shows higher variation which makes it moderately sensitive to training data.

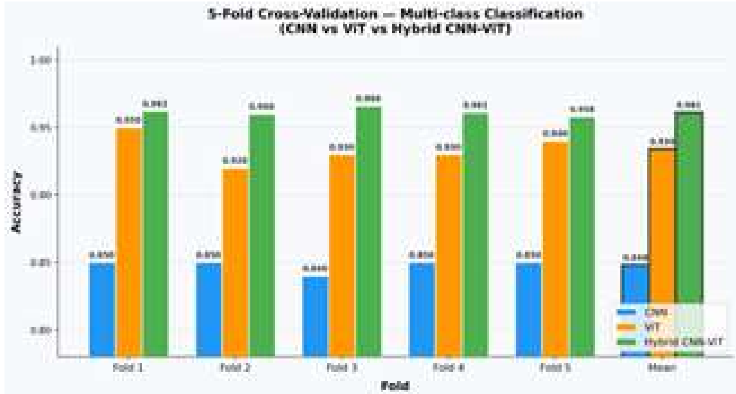


Fig. 14. *5-fold CV bar chart: Multi-class classification. The Hybrid achieves the highest mean (0.961) with the lowest variance (±0.003),confirming table superior performance.*

Table 5. 5-Fold Cross-Validation: Multi-class Classification.

Fold No.	CNN	ViT	Hybrid CNN-ViT
1	0.850	0.950	0.962
2	0.850	0.920	0.960
3	0.840	0.930	0.966
4	0.850	0.930	0.961
5	0.850	0.940	0.958

B) Hierarchical Cross-Validation:

The Hybrid CNN-ViT model outperforms all other models in hierarchical classification by achieving a mean accuracy of 0.954 which demonstrates its effectiveness at multi-stage classification .The results demonstrate that the hybrid model benefits from its approach to extract both local and global features.

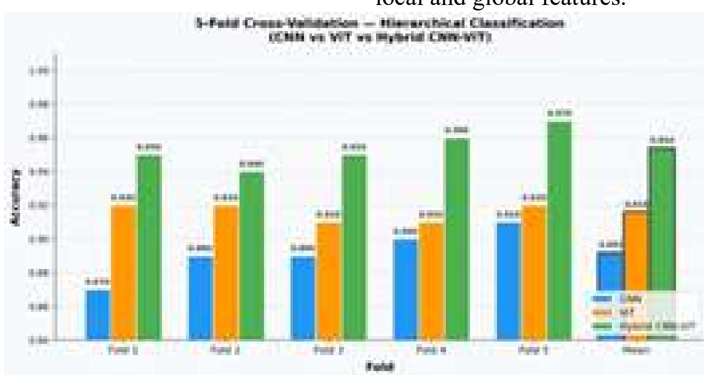


Fig.15 .5*-fold CV bar chart: Hierarchical classification. The Hybrid out performs both stand alone models in every fold, achieving mean 0.954.*

Table 6. 5*-Fold Cross-Validation: Multi-class Classification.*

Fold No.	CNN	ViT	Hybrid CNN-ViT
1	0.870	0.920	0.950
2	0.890	0.920	0.940
3	0.890	0.910	0.950
4	0.900	0.910	0.960
5	0.910	0.920	0.970

C) Binary Cross-Validation:

The binary (anomaly) classification models show exceptional results because all their systems reach performance levels above 99.7percent accuracy. The model differences show minimal variation because the standard deviation values are low, which demonstrates that all models perform consistently across different testing scenarios. The results demonstrate that binary classification represents an easier task because local features enable successful discrimination.

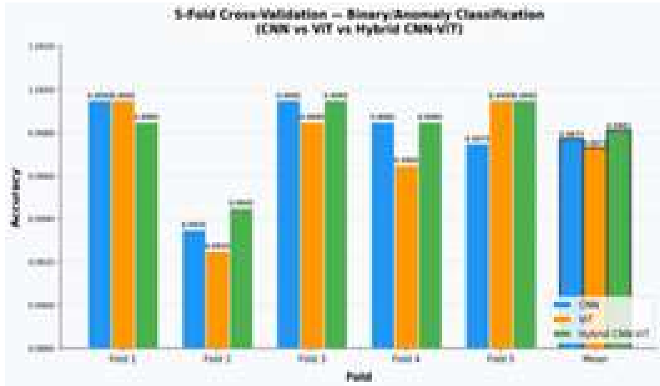


Fig. 16.*5-fold CV bar chart: Binary classification.*

Table 6. 5-Fold Cross-Validation: Binary/Anomaly Classification.

Fold No.	CNN	ViT	Hybrid CNN-ViT
1	0.9995	0.9995	0.9985
2	0..9935	0.9925	0.9945
3	.0.9995	0.9985	0.9995
4	0.9985	0..9965	0.9985
5	0.9975	0..9995	0.9995

4.6 Evaluation Metrics

The assessment of model performance depends on three testing methods which include Test Accuracy as the main measurement together with Confusion Matrix which shows how different classes perform and reveals repeated error patterns and ROC Curves with AUC which assesses how well the system distinguishes between two classes and detects anomalies and Macro-averaged F1-Score which provides equal assessment of all classes irrespective of their occurrence rate.

4.7 Hypothesis Testing

The paired t-test is used to analyze data from multiple experimental trials which produced repeated measurements of the same variables. The test is used to check the data to confirm its validity. The paired design enables a direct comparison between corresponding performance measures of baseline and Hybrid models, thereby minimizing inter-sample variability. The testing framework establishes that all observed differences result from actual model performance which exists independently of random changes. The statistical methods used in this research validate the proposed method through rigorous and objective assessment. The experimental results showed that all test runs maintained consistent performance because no unusual changes appeared during the testing process. The test results show consistent statistical performance because there were no unexpected variations during the testing process which confirmed stable results across all testing conditions.

The hypotheses are defined as follows:

Null Hypothesis (H_0):

$$H_0 : \mu = 0 \quad \textbf{(1)}$$

Both models perform equally.

Alternative Hypothesis (H_1):

$$H_1 : \mu > 0 \quad \textbf{(2)}$$

Hybrid model performs significantly better.

Decision Rule:

If calculated $t > t_{\alpha,df}$ at significance level $\alpha = 0.05$, reject H0 and conclude that the Hybrid model performs significantly better than the baseline CNN or ViT model.

Test Statistic: The paired t-test statistic is computed as

$$t = \frac{\bar{d}}{sd / \sqrt{n}} \quad (3)$$

where $\bar{d}$ is the mean of the differences between paired observations, sd is the standard deviation of the differences, and n is the number of samples. The statistical protocol was executed by comparing CNN and ViT against the Hybrid model across both hierarchical and multiclass classification tasks. The obtained results are as follows:

Hierarchical Classification

- *CNN vs .Hybrid:*

T-Statistic: -12.655; P-Value: 2.24×10^{-4} => Reject H_0 ($p<0.05$).

- *ViT vs. Hybrid:*

T-Statistic: -6.5169;P-alue: 2.86×10^{-3}=>Reject H_0 ($p<0.05$).

Multiclass Classification

- *CNN vs. Hybrid:*

T-Statistic: -34.7851;P-Value: 4.08×10^{-6}=>Reject H_0 ($p<0.05$).

- *ViT vs. Hybrid:*

T-Statistic: -4.0000;P-alue: 1.61×10^{-2}=>Reject H_0 ($p<0.05$).

Since all p-values are less than the significance level (α=0.05),the null hypothesis rejected in all cases. The negative t-statistics indicate that the Hybrid model consistently achieves superior performance compared to the baseline models. The Hybrid model showed performance improvements which the hypothesis testing proved to be statistically significant because they did not result from random variation, thus proving the method's effectiveness and reliability.

The testing results show that CNN, ViT, and the Hybrid model perform at the same level throughout all tests for measuring anomalies. The 5-fold cross-validation bar chart for binary classification in Figure 16 shows that all models achieve nearly perfect accuracy because Area under the Curve values reach approximately 0.99 while there exist no significant differences between the five test groups. The system demonstrates consistent performance because it detects anomalies with the same effectiveness across different architectural designs. The result shows that all models can effectively handle anomaly detection which does not limit the system. The test results show no important difference between the models because all models maintain the same capability to detect anomalies while the Hybrid model still offers additional boosts to various performance measurements beyond binary classification.

The null hypothesis states that both models show identical performance because their respective mean values remain at zero. The alternate hypothesis states that the Hybrid model shows superior performance because its mean value exceeds zero. Among all three models, the Hybrid model delivers the best accuracy results across every classification test. The standalone CNN and ViT models demonstrate similar accuracy results because both models use the same experimental conditions to generate their results. The Hybrid model establishes its multi-branch fusion design because it needs to achieve larger feature space extraction which neither of its individual components can deliver.

5. CONCLUSION

The researcher created a complete deep learning system for plant disease detection through their research by combining three different classification methods which include binary classification and multiclass classification and hierarchical classification methods. The project develops a multi-model system through its combination of CNN and Vision Transformer and hybrid architectural systems which enable it to meet actual classification needs. The system uses multiple classification methods which allow it to do anomaly detection and complete category identification and organized decision-making through a single operational system.

The experimental analysis shows that the proposed system can handle classification tasks which vary in complexity. The CNN models capture local spatial information while Vision Transformers create global relationships and hybrid models use both methods to improve feature representation. The hierarchical classification system develops better interpretability through its implementation of multi-stage classification systems. The model training process for 30 and 50 epochs enables researchers to understand how learning occurs and when models reach their final state. The proposed framework delivers an adaptable solution which can grow in capacity while using a systematic approach for detecting plant diseases. The research develops practical agricultural systems through its multiple classification approaches which extend beyond standard methods.

6. REFERENCES

[1] Upadhyay A, Chandel NS, Singh KP, Chakraborty SK, Nandede BM, Kumar M, Subeesh A, Upendar K, Salem A, Elbeltagi A. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. Artificial Intelligence Review. 2025 Jan 17;58(3):92.

[2] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

[3] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.

[4] Jawed MM, Tufail FA, Ahmed MZ, R AS, Nallusamy P, Raja KT. A hybrid deep learning framework using convolutional and transformer models for robust plant disease classification. Scientific Reports. 2026 Feb 18;16(1):9704.

[5] Hughes D, Salathé M. An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060. 2015 Nov 25.

[6] Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Computers and electronics in agriculture. 2018 Feb 1;145:311-8.

[7] Abdal MN, Islam K, Oshie MH, Haque MA. A CNN based model for plant disease classification using transfer learning. In2023 26th International

Conference on Computer and Information Technology (ICCIT) 2023 Dec 13 (pp. 1-6). IEEE.

[8] Sahu P, Sinha VK. Plant disease detection using transfer learning with DL model. InMobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2022 2022 Jul 16 (pp. 169-180). Singapore: Springer Nature Singapore..

[9] M. M. Sreedhar, K. Mamatha, Madhu Babu Mallepalli "Deep learning-assisted framework for plant disease identification in CNN models", International Journal of Engineering, Science and Advanced Technology, 25(12):467-474, December 2025.

[10] Ali M, Salma M, El Haji M, Jamal B. Plant disease detection using vision transformers. International Journal of Electrical and Computer Engineering (IJECE). 2025 Apr;15(2):2334-44.

[11] Vallabhajosyula S, Sistla V, Kolli VK. A novel hierarchical framework for plant leaf disease detection using residual vision transformer. Heliyon. 2024 May 15;10(9).

[12] Durgadevi P, Murikinati HR, Zagabathuni M. Tomato Leaves Disease Classification Using Vision. InProceedings of the International Conference on Intelligent Systems and Digital Transformation (ICISD 2025) 2025 Oct 30 (p. 308). Springer Nature.

[13] Dutta P, Sathi KA, Hossain MA, Dewan MA. Conv-ViT: a convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. Journal of Imaging. 2023 Jul 10;9(7):140.

[14] Swamy VK, Anusha R, Begum SF. HyTran-CottonNet: A Hybrid Vision Transformer-CNN Architecture with Cross-Disease Attention for Robust Cotton Leaf Disease and Severity Classification. In2025 IEEE North Karnataka Subsection Flagship International Conference (NKCon) 2025 Sep 27 (pp. 1-6). IEEE.

[15] Li YC, Li R, Zhu JB. Hybrid Model based on CNN-Transformer for Tomato Pest and Disease Identification. IAENG International Journal of Computer Science. 2025 Nov 1;52(11).

[16] Shahi TB, Xu CY, Neupane A, Guo W. Recent advances in crop disease detection using UAV and deep learning techniques. Remote Sensing. 2023 May 6;15(9):2450.

[17] Zeng J, Jia B, Song C, Ge H, Shi L, Kang B. CDPNet: a deformable ProtoPNet for interpretable wheat leaf disease identification. Frontiers in Plant Science. 2025 Nov 6;16:1676798.

[18] Oad A, Koondhar IH, Dong F, Liu W, Zou B, Liu W, Chen Y, Wu Y. Symmetry-Aware SwinUNet with Integrated Attention for Transformer-Based Segmentation of Thyroid Ultrasound Images. Symmetry. 2026 Jan 10;18(1):141.

[19] Voetman R, Dijkstra W, Wolters JE, Dijkstra K. Identification of potato virus y in potato plants using deep learning and gradcam verification. InProceedings of SAI Intelligent Systems Conference 2023 Sep 7 (pp. 223-244). Cham: Springer Nature Switzerland.

[20] Hoque MS, Biswas S, Sakib MS, Leander R, Umbaugh SE. Hybrid image enhancement for thermographic imaging in canine bone cancer detection. InThermosense: Thermal Infrared Applications XLVII 2025 May 29 (Vol. 13470, pp. 156-164). SPIE.

[21] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[22] Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4. No. 4., springer, 2006.