AN ENSEMBLE-BASED DEEP LEARNING APPROACH USING CONVNEXTV2 AND BEIT MODELS FOR WATERMELON DISEASE CLASSIFICATION

Tuong Le

Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam

lc.tuong@hutech.edu.vn

ABSTRACT — This study proposes MelonDx, a novel ensemble-based framework for early classification of watermelon diseases, aimed at overcoming the limitations and subjectivity of traditional diagnostic methods through the application of advanced deep learning models. MelonDx integrates ConvNeXtV2, a state-of-the-art convolutional neural network, with BEiT, a transformer encoder model, via ensemble learning to harness their complementary strengths in feature extraction and disease classification. The framework was rigorously evaluated on a dedicated watermelon disease dataset using metrics such as accuracy, F1 score, and AUC, and demonstrated superior performance with results of 99.43% accuracy, 99.43% F1 score, and 99.99% AUC—outperforming both individual models and other state-of-the-art approaches. By combining the advantages of CNN and transformer-based architectures, MelonDx enhances accuracy and robustness, offering a promising solution for precision agriculture that can help reduce crop losses and optimize production efficiency through accurate and timely disease detection.

Keywords— Ensemble Learning, Deep Learning, Watermelon Disease Classification.

I. INTRODUCTION

Early detection of plant diseases using deep learning is an important component of modern agriculture, as it can significantly reduce crop losses and improve food security [[1]]. Among various agricultural crops, watermelon is susceptible to a number of diseases that can adversely affect yield. Therefore, timely and accurate detection of watermelon diseases is essential for effective management and control of these diseases. Traditional disease identification methods often involve manual inspection by experts, which is both time-consuming and prone to error due to dependence on the experience of experts, especially when dealing with large-scale crops. Therefore, there is a need to develop more efficient and automated methods for detecting watermelon diseases using the power of artificial intelligence to assist farmers in crop management.

In recent years, deep learning-based image classification has demonstrated significant potential in many fields, including healthcare [[2]-[3]], robotics [[4]], and smart agriculture [[4]-[20]]. In agriculture, deep learning has also opened up a new, time-efficient, accurate, and labor-saving approach to crop management and disease detection. Sinha and Dhanalakshmi [[4]] provided a comprehensive overview of the transformative role of IoT in modernizing agriculture, emphasizing automation, data analytics, and minimal human intervention to enhance productivity and efficiency. In addition, recent studies have furthered this vision by collecting the agricultural datasets and disease detection systems. For example, the Chilli and Onion Leaf Dataset (COLD) was introduced by Aishwarya and Reddy [[6]] for identifying diseases like purple spot and nutrient deficiencies, while a bean leaf image dataset for diagnosing bean rust and anthracnose was created by Laizer et al. [[7]], providing significant benefits to smallholder farmers. Subsequently, the strong developments of machine learning, especially deep learning, have played a key role in the transformation of modern agriculture, especially the problem of early disease detection of agricultural crops. For instance, high accuracy in plant disease detection is reported through machine vision employing MobileNet [[8]], SSD768 [[9]], and VGG-16 [[10]] models. Furthermore, Santoso et al. [[11]] conducted a comparative analysis of convolutional neural networks and DenseNet121 utilizing transfer learning approaches in agriculture, specifically focusing on crop leaf disease identification. Their findings revealed the superior performance of the DenseNet121 model over the CNN model in classifying diseases on tomato leaves. These studies highlighted the potential of artificial intelligence, particularly machine learning, to enhance crop management and minimize agricultural losses.

Recently, two of the latest pre-trained deep learning models, Convolutional Neural Networks (CNNs) and Transformer-based models, such as ConvNeXtV2 [[12]] and BEiT [[13]], have shown good performance in many image classification tasks and are widely applied to various problems [[14]-[16]]. To begin with, in 2024, an unsupervised clustering algorithm based on BEiT and feature extraction [[14]] is proposed for wafer pattern classification, aiming to improve process analysis and optimization while eliminating the need for manual labeling. The results of this study demonstrate a significant performance improvement, with a 16.2 percentage point increase in the Jaccard Coefficient (JC) and a 14.4 percentage point improvement in the Fowlkes-Mallows (FM) index compared to the BEiT model. Following this, a novel lightweight and mobile-friendly hybrid deep learning model for skin cancer detection, combining ConvNeXtV2 blocks and focal self-attention mechanisms to improve diagnostic accuracy [[15]] was developed by Ozdemir and Pacal. The proposed model achieved 93.60%

accuracy, 91.69% precision, 90.05% recall, and a 90.73% F1 score on the ISIC 2019 dataset, outperforming baseline models such as ResNet50 and Swinv2-Base by 10.8% and 3.3% in accuracy, respectively. Then, machine learning and signal processing are combined to enhance the early detection of Parkinson's disease, a neurodegenerative disorder affecting motor functions [[16]]. The BEiT Transformer model achieved the highest accuracy of 90%, outperforming SVM, MLP, and XGBoost, highlighting the potential of deep learning in improving Parkinson's disease diagnosis.

In recent years, ensemble learning has gained significant attention for its ability to enhance classification accuracy by combining multiple models. While each model demonstrates high performance on specific datasets, integrating them within an ensemble framework can further improve overall accuracy by leveraging their complementary strengths. An ensemble framework [[17]] integrates multiple models to enhance classification accuracy by harnessing the strengths of each individual model, leading to superior performance compared to any single model. As a result, this approach has been widely adopted in the research community for its effectiveness in improving both model robustness and reliability. For example, an ensemble-based deep reinforcement learning method for vehicle routing problems (VRPs) under distribution shift [[18]] was developed, enhancing generalization by learning diverse sub-policies. This method outperforms state-of-the-art neural models on random instances with various distributions and generalizes effectively on benchmark datasets from TSPLib and CVRPLib. Similarly, Jiang et al. [[19]] introduced an ensemble-based deep learning approach for detecting and classifying submerged arc weld defects during Non-Destructive Testing (NDT). The model achieves an impressive accuracy of 93.12% in fault detection and classification, outperforming existing methods and demonstrating significant potential for real-time industrial applications. Overall, ensemble learning continues to show great promise in advancing various domains by improving both performance and generalization capabilities.

In 2024, the Watermelon dataset (we denote it as the WMelon dataset from this point until the end of the manuscript to avoid confusion) was collected by Nakib and Mridha [[20]], which consists of images of healthy watermelons and those affected by three common diseases in watermelon plants: Mosaic Virus, Anthracnose, and Downy Mildew. By leveraging CNN and DenseNet121 models, they achieved remarkable disease recognition accuracies of 96% and 98%, respectively. These results demonstrate the effectiveness of these deep learning models in accurately detecting diseases in watermelon.

However, further improvements in accuracy are possible by leveraging more advanced deep learning models and ensemble learning techniques. To address this, our study proposes an ensemble-based deep learning framework, MelonDx, for watermelon disease classification for the WMelon dataset. MelonDx integrates ConvNeXtV2, a state-of-the-art CNN model, with BEiT, a transformer-based model known for its strong performance in visual tasks, through an ensemble approach. As the latest and most effective architectures for image classification, ConvNeXtV2 and BEiT enhance accuracy and robustness, ensuring reliable disease detection in watermelon plants. By combining the power of these advanced models, MelonDx aims to improve performance in watermelon disease classification. This method can significantly improve the accuracy of watermelon disease detection, contributing to more efficient agriculture.

II. MATERIALS AND METHODS

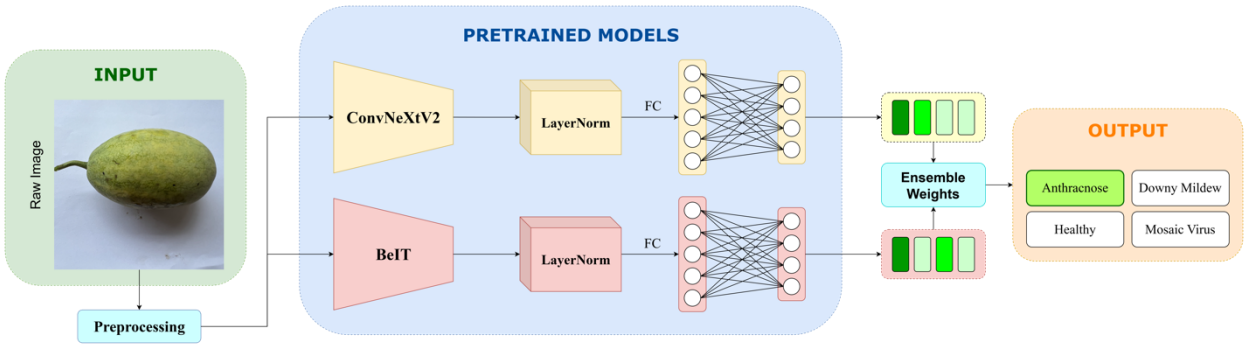


Figure 1. *Overview of the MelonDx framework.*

This section introduces MelonDx (Figure 1), short for "Melon Diagnosis", for watermelon disease diagnosis on the WMelon dataset. MelonDx is an ensemble-based deep learning framework carefully designed to achieve high accuracy and efficiency in watermelon disease identification, addressing a critical need in agricultural disease management.

A. DATASET OVERVIEW AND PREPROCESSING

Table 1. *Overview of watermelon diseases on WMelon dataset.*

Category	Description	Illustration
Mosaic Virus	The watermelon mosaic virus is a potyvirus that causes symptoms such as leaf chlorosis with a yellow or green mosaic pattern, stunted growth, deformed leaves, and reduced fruit yield.	
Downy Mildew	Downy mildew on watermelon plants, caused by Pseudoperonospora cubensis, results in leaf yellowing, angular lesions, and greyish downy growth beneath the leaves, and can be controlled through resistant varieties, proper spacing, reduced overhead watering, fungicides, and regular inspections.	
Anthracnose	Anthracnose, caused by Colletotrichum orbiculare, affects watermelon plants by causing lesions on leaves, vines, and fruit, and can be managed through the use of disease-free seeds, crop rotation, clean cultivation, and appropriate fungicides.	
Healthy	A healthy watermelon leaf is characterized by a bright green color, smooth texture, turgidity, and the absence of yellowing or brown necrotic patches, with consistent growth and no visible damage or pests.	

In 2024, the WMelon dataset, a comprehensive collection specifically designed for watermelon disease classification, was introduced [[20]]. This dataset includes four distinct categories: healthy watermelons and those affected by three common diseases, including Mosaic Virus, Anthracnose, and Downy Mildew. Table 1 provides an overview of watermelon diseases with corresponding images in the WMelon dataset. This dataset was collected in collaboration with agricultural experts, thus ensuring high accuracy and reliability. The images were collected on June 25, 2023, at the Regional Horticultural Research Station in Lebukhali, Patuakhali and were captured under different lighting and environmental conditions to represent the real-life environment. This dataset is an important foundation for developing machine vision-based AI models to support accurate diagnosis and prediction of watermelon diseases.

The WMelon dataset consists of 1,155 images categorized into four classes: Mosaic Virus, Downy Mildew, Anthracnose, and Healthy. To ensure consistency in evaluating models, the dataset was divided into three subsets: Training Set (70% of the total data, including 807 images), Validation Set (15% of the total data, including 172 images), and Test Set (15% of the total data, including 176 images). The detailed distribution of classes across the above three subsets is presented in Table 2. The Training Set will be used to train the model, while the Validation Set is used to observe and optimize the parameters of the proposed model. Finally, the Test Set is used to evaluate the performance of the experimental models, with three key metrics: Accuracy, F1 score, and AUC, providing a comprehensive assessment of model effectiveness.

To ensure the consistency of the input images, we process the images through three main pre-processing steps:

- Image conversion. All images in the WMelon dataset are normalized to RGB format for uniform color representation.
- Image Resizing. In this step, each image is resized to 224×224 pixels to fit the model's input requirements.

- Pixel Normalization. This study normalizes the pixel values of all input images to improve the stability of the training process and enhance the convergence of the model.

The data split and preprocessing steps outlined above ensure standardized, high-quality data, making it ideal for training and evaluating deep learning models in watermelon disease classification, thus enhancing model performance and reliability.

Table 2. *The WMelon dataset composition across categories and data splits.*

Category	WMelon dataset	Train Set	Validation Set	Test Set
Mosaic Virus	415	290	62	63
Downy Mildew	380	266	57	57
Anthracnose	155	108	23	24
Healthy	205	143	30	32
Total	1,155	807	172	176

B. PRE-TRAINED MODELS

Pre-trained models have become the foundation of modern machine learning, providing a powerful starting point for a variety of tasks by leveraging knowledge gained from large-scale datasets. Leveraging these models as a foundation will reduce training time, speed up convergence, and significantly improve generalization to subsequent tasks. In this section, we focus on two notable pre-trained models, ConvNeXtV2 [[12]] and BEiT [[13]], which have played an important role in improving performance for recent image-based applications. MelonDx incorporates these two models as integral components of its framework, leveraging the strengths of ensemble techniques to enhance both accuracy and efficiency in detecting watermelon diseases on the WMelon dataset.

ConvNeXtV2. This model was introduced in 2023 by Woo et al. [[12]]. In this study, the authors developed a fully convolutional masked autoencoder (FCMAE) framework and introduced a new layer called Global Response Normalization (GRN) into the ConvNeXt architecture.

The FCMAE framework, shown in Figure 2A, has an asymmetric autoencoder architecture consisting of a sparse convolution-based ConvNeXt encoder and a lightweight ConvNeXt block decoder. The encoder operates exclusively on visible pixels, while the decoder reconstructs the image using encoded pixels and mask tokens. The loss function is computed exclusively on masked regions, ensuring efficient and targeted learning.

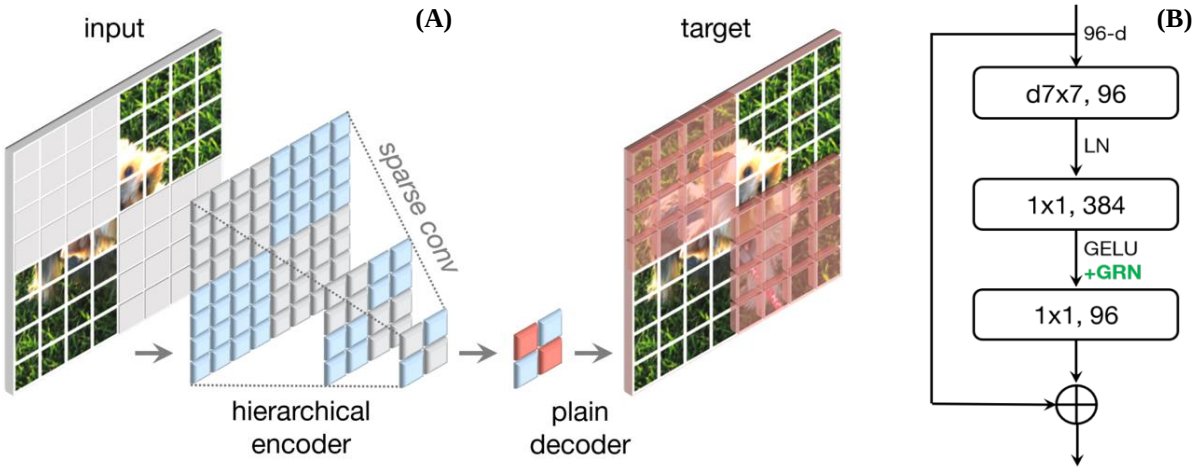


Figure 2. *The FCMAE framework (A); and Global Response Normalization (GRN) (B) [[12]].*

This study also introduced the GRN technique to make FCMAE pretraining more effective in conjunction with the ConvNeXt architecture. Given an input feature, $X \in R^{H\times W\times C}$, the proposed GRN unit consists of three steps:

- **Global feature aggregation.** This study firstly aggregates a spatial feature map X_i into a vector gx with a global function $G(\cdot)$:

$$G(X) := X \in R^{H\times W\times C} \rightarrow gx \in R^{C}. \quad (1)$$

- **Feature normalization.** Next, a response normalization function $N(\cdot)$ was applied to the aggregated values as follows:

$$N(\|X_i\|) := \|X_i\| \in \mathrm{R} \rightarrow \frac{\|X_i\|}{\sum_{j=1,\ldots,C}\|X_j\|} \in \mathrm{R}, \tag{2}$$

where $\|X_i\|$ is the L2-norm of the i-th channel.

- **Feature calibration.** Finally, GRN calibrates the original input responses using the computed feature normalization scores:

$$X_i = X_i * N(G(X)_i) \in R^{H \times W}. \tag{3}$$

The ConvNeXtV2 model incorporates the GRN layer into the original ConvNeXt block, as illustrated in Figure 2B, to enhance inter-channel feature competition. This model utilizes pre-trained weights from the ConvNeXtV2-nano model trained on the ImageNet-1K dataset (convnextv2_nano.fcmae_ft_in1k from the TIMM library). We freeze all layers except the final stage and partially unfreeze selected blocks from the penultimate stage to balance performance and computational efficiency.

BEiT. The Bidirectional Encoder Representation from Image Transformers (denoted by BEiT) was introduced by Bao et al. [[13]] and pre-trained on the large-scale ImageNet-21k dataset using a self-supervised learning approach at a resolution of 224×224 pixels in 2022. BEiT is inspired by BERT [[21]], a pioneering model in natural language processing, and adopts a similar pretraining objective known as Masked Image Modeling (MIM). This task utilizes two representations for each image: image patches and visual tokens (Figure 3).

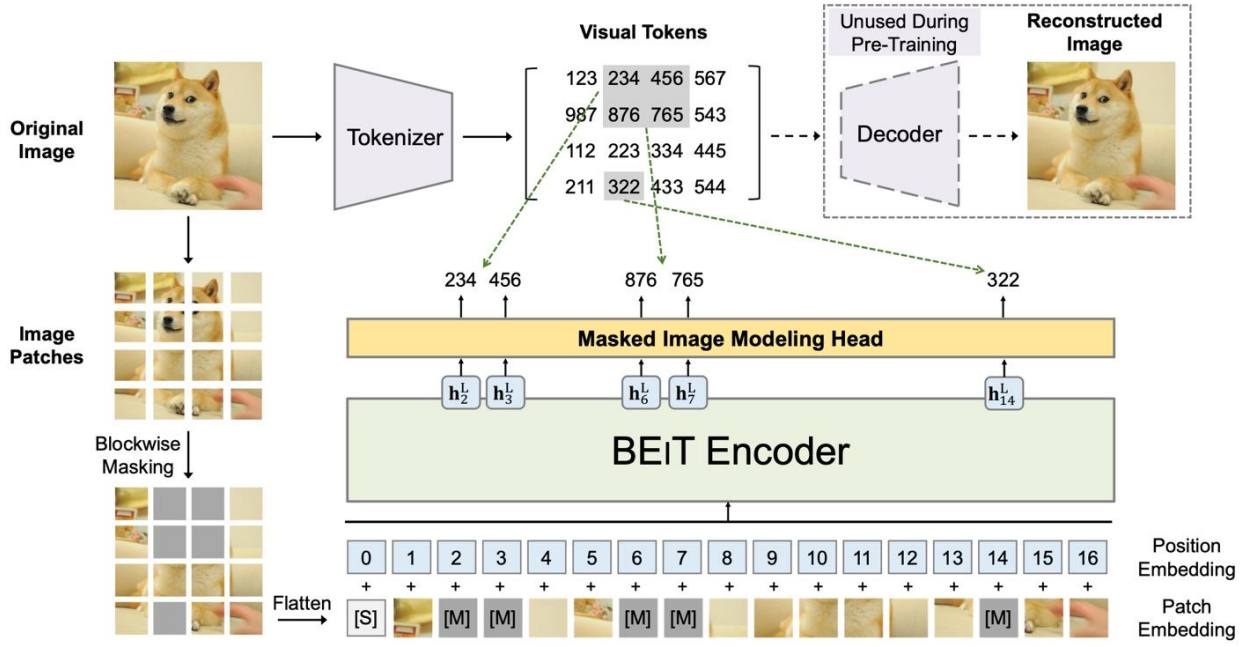


Figure 3. *Overview of BEiT model [[13]].*

The pretraining process of BEiT involves several key steps. First, an image is divided into a grid of patches, which serve as input representations for the backbone transformer. Second, BEiT "tokenizes" the image into discrete visual tokens, generated using the latent codes of a discrete variational autoencoder (dVAE) [[22]]. Before pretraining, BEiT learns an "image tokenizer" through an autoencoding-style reconstruction process, where images are converted into discrete visual tokens based on a learned vocabulary.

During pretraining, BEiT leverages two views of each image: image patches and visual tokens. A random subset of image patches is masked (represented as gray patches in Figure 3) and replaced with a special mask embedding [M]. These masked patches are then input into the vision Transformer backbone. The pretraining objective is to predict the visual tokens of the original image using the encoded representations of the corrupted image.

Output Layer Modification. To adapt the pre-trained models to the task of watermelon disease classification in our task, the original output layers of both pre-trained models (i.e., ConvNeXtV2 and BEiT) were replaced with fully connected layers designed to output four distinct labels, corresponding to our target classes (see Figure 1).

This modification was imperative to ensure that the models could effectively discriminate between the classes pertinent to the WMelon dataset.

The tuning process consists of two main components: a normalization layer and a classification layer. Firstly, the existing normalization layer is replaced with LayerNorm configured as LayerNorm((768,), eps=1e-12, elementwise_affine=True). This layer normalizes the inputs across the features of each individual data sample, which stabilizes the learning process and speeds up convergence. The parameters eps=1e-12 and elementwise_affine=True ensure numerical stability and allow the layer to learn affine transformation parameters, respectively. This normalization process is important, especially when dealing with different input distributions, as it reduces the internal drift of the dependent variable and enhances the model's generalizability across different data samples.

Secondly, the classifier component was modified by replacing the original output layer with a fully connected Linear layer defined as Linear(in_features=768, out_features=4, bias=True). The in_features parameter corresponds to the dimensionality of the output of the previous layer (768 features), while out_features = 4 matches with the number of target classes in our classification task. The parameter bias (set to bias=True) allows the model to fit the data more efficiently by introducing an extra degree of freedom during training. This layer serves as the final decision-making component, aggregating the learned representations and mapping them to the desired output layers. This configuration ensures that the model's capabilities are tailored to the complexity and requirements of the WMelon dataset, improving performance and accuracy in classification tasks.

C. SOFT VOTING ENSEMBLE

To improve the classification performance, an ensemble learning strategy was implemented, combining predictions from the ConvNeXtV2 and BEiT models. This approach exploits the complementary strengths of both models to enhance the overall accuracy and robustness in watermelon disease classification. The ensemble probability for a given class is calculated using Equation (4) as follows:

$$P_{ensemble} = (1-\alpha)P_{ConvNeXtV2} + \alpha P_{BEiT} \quad (4)$$

where:

- $P_{ConvNeXtV2}$ represents the probability predicted by the ConvNeXtV2 model for a specific class.
- P_{BEiT} represents the probability predicted by the BEiT model for the same class.
- $P_{ensemble}$ is the final ensemble probability.
- α is a weighting parameter that controls the contribution of each model.

The model's prediction result is determined by selecting the class with the highest $P_{ensemble}$. This ensemble approach effectively leverages the power of ConvNeXtV2 and BEiT, combining their predictive capabilities to achieve better performance and reduce the inherent bias of each model. In this study, after thorough observation and analysis, we identified that $\alpha = 0.2$ yielded the best classification results for the WMelon dataset. Therefore, in this research, we used $\alpha = 0.2$ for our experiments. This highlights the potential of advanced architectures and ensemble strategies for robust and accurate disease classification in agricultural applications.

III. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETTING

Table 3. *Training configuration for ConvNeXtV2 and BEiT models.*

No	Parameter	Value	Description
1	Batch Size	16	Optimized for computational efficiency.
2	Learning Rate	1×10^{-5}	Cosine scheduler used for gradual reduction.
3	Weight Decay	0.095	Configured to mitigate overfitting.
4	Mixed Precision Training	Enabled (FP16)	Enhanced computational speed and reduced memory usage.

Experiments are conducted to evaluate the performance of the proposed MelonDx framework, ConvNeXtV2, BEiT, and two existing state-of-the-art methods (CNN and DenseNet121) in disease detection on the WMelon dataset.

All experiments are performed on high-performance machines equipped with NVIDIA T4×2 GPUs, using CUDA to accelerate computation and training.

The proposed model is implemented using the Hugging Face Transformers library and the PyTorch framework for model development and training. Data preprocessing is done using NumPy and Pandas. The key parameters of ConvNeXtV2 and BEiT models, including batch size, learning rate, weight decay, and mixed precision training, are detailed in Table 3.

B. EVALUATION METRICS

The experimental models were evaluated using three common metrics including Accuracy, which measures the overall accuracy of the model's predictions; F1 score, which balances precision and recall, providing insight into the model's performance on imbalanced datasets; and AUC (Area Under the Receiver Operating Characteristic Curve), which assesses the model's ability to distinguish between classes.

- **Accuracy** measures the overall correctness of the model by calculating the ratio of correct predictions (True Positives and True Negatives) to the total number of predictions. It is given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

  where TP (True Positives) are correctly predicted positives, TN (True Negatives) are correctly predicted negatives, FP (False Positives) are negatives incorrectly predicted as positives, and FN (False Negatives) are positives incorrectly predicted as negatives.

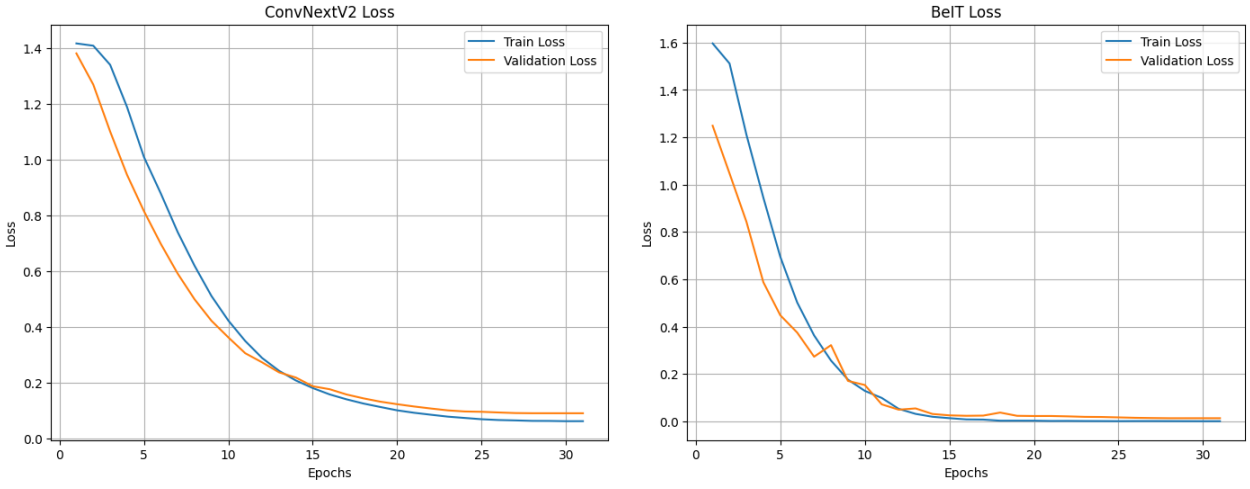
- **F1 score** is the harmonic mean of Precision and Recall, providing a balance between the two. It is especially useful when dealing with imbalanced datasets:

$$F1\ score\ = 2 \times \frac{\mathrm{Pr}\,e\,cision\ \times Recall}{\mathrm{Pr}\,e\,cision\ +\ Recall} \tag{6}$$

  where $Precision\ = \frac{TP}{TP+FP}$ and $Recall\ = \frac{TP}{TP+FN}$.

- **Area Under the Receiver Operating Characteristic (ROC) Curve** (denoted by AUC) evaluates the model's ability to distinguish between positive and negative classes. It is not directly derived from the confusion matrix but instead from the model's predicted probabilities. To compute AUC, the model first generates predicted probabilities for each sample belonging to the positive class. Then, a ROC curve is plotted by varying the classification threshold and calculating TPR and FPR at each threshold. AUC is the area under the ROC curve, and a value closer to 1 indicates that the model distinguishes between the classes more effectively.

C. OPTIMIZING THE NUMBER OF EPOCHS IN MODEL TRAINING



(A) (B)

Figure 4. *Loss curve analysis of ConvNeXtV2 and BEiT on the WMelon dataset.*

This experiment aims to optimize the number of epochs during model training, with the objective of identifying the ideal epoch count that ensures effective model learning while avoiding overfitting, ultimately enhancing

training efficiency. Figure 4 presents a detailed loss curve analysis for the ConvNeXtV2 and BEiT models, showcasing their performance on the WMelon dataset.

As shown in Figure 4, both models demonstrate no obvious signs of overfitting. ConvNeXtV2 achieves optimal performance at 25 epochs (Figure 4A), while BEiT reaches its peak performance at 15 epochs (Figure 4B). Based on these observations, we choose 25 epochs for ConvNeXtV2 and 15 epochs for BEiT to be implemented within the MelonDx framework.

D. EXPERIMENTAL RESULTS

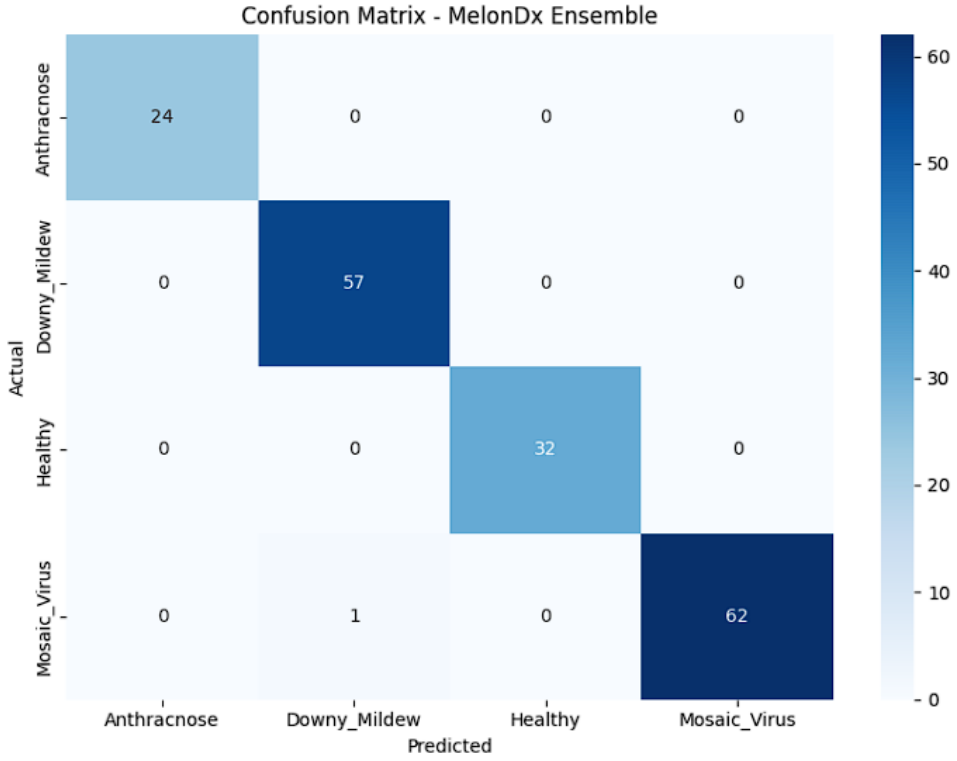


Figure 5. *Confusion matrix for evaluating the MelonDx framework.*

In this section, we first present the confusion matrix for the MelonDx framework, followed by the calculation of key metrics, including Accuracy, F1 score, and AUC, derived from the confusion matrix. The performances of the experimental methods, including ConvNeXtV2, BEiT and the MelonDx framework on these metrics are then compared with the results of CNN and DenseNet121 models from [[20]], demonstrating the enhanced performance and robustness of the MelonDx framework.

According to the results presented in Figure 5, MelonDx framework misclassified only five images: one image of Healthy and four images of Downy Mildew, which were incorrectly predicted as Mosaic Virus. These results demonstrate the model's high classification accuracy, with only a small number of misclassifications occurring between disease categories.

Table 4. *Performance comparison of experimental models for watermelon disease classification on the WMelon dataset.*

No	Approach	Accuracy (%)	F1 score (%)	AUC (%)
1	CNN [20]	95.67	95.69	99.92
2	DenseNet121 [20]	98.27	98.27	99.81
3	ConvNeXtV2	98.30	98.29	99.98
4	BEiT	98.86	98.86	99.90
5	MelonDx (The proposed model)	99.43	99.43	99.99

Based on the Confusion Matrix, Table 4 presents the comparative results of the performance of different experimental models used for watermelon disease detection on the WMelon dataset. Among the experimental

models, the MelonDx framework achieved the highest accuracy at 99.43%, an F1 score of 99.43%, and an impressive AUC of 99.99%, demonstrating superior performance. Meanwhile, ConvNeXtV2 and BEiT also demonstrate strong results across all metrics, especially in terms of accuracy and AUC. Similarly, CNN and DenseNet121 methods achieved accuracy rates of 95.67% and 98.27%, respectively, along with F1 scores of 95.69% and 98.27%, and AUC values of 99.92% and 99.81%. Although DenseNet121 and BEiT also showed promising results, the MelonDx framework stood out, highlighting the effectiveness of the ensemble-based approach in improving watermelon disease classification accuracy on the WMelon dataset.

IV. CONCLUSION

This study developed an ensemble-based deep learning framework (MelonDx) for watermelon disease classification, leveraging the strengths of ConvNeXtV2, a highly efficient CNN model, and BEiT, a transformer-based model excelling at capturing long-range dependencies and spatial features in images. MelonDx effectively addresses the problem of watermelon disease classification on the WMelon dataset. Specifically, the experimental results demonstrate the superiority of the proposed approach, achieving an impressive accuracy of 99.43%, an F1 score of 99.43%, and an AUC of 99.99%, outperforming individual models (i.e., ConvNeXtV2 and BEiT) and existing state-of-the-art methods (CNN and DenseNet121 models) on the WMelon dataset. These results demonstrate the potential of applying ensemble learning methods to improve disease prediction in watermelon plants and thus contribute to precision agriculture.

For future research, we plan to collect a new dataset from Vietnam, including many different samples from different seasons and regions, which can improve the generalizability of the model to local conditions. Additionally, integrating multimodal data, such as environmental and sensor data, with input image data, can provide additional context for disease detection for watermelon plants. Finally, optimizing lightweight deep learning models and employing advanced optimization techniques will facilitate real-time deployment on disease recognition devices for practical on-farm applications.

V. REFERENCE

[1] Pandey, D. K., & Mishra, R. (2024). Towards sustainable agriculture: Harnessing AI for global food security. *Artificial Intelligence in Agriculture*, 12, 72-84.

[2] Vo, M. T., Vo, A. H., & Le, T. (2022). A robust framework for shoulder implant X-ray image classification. *Data Technologies and Applications, 56(3),* 447–460.

[3] Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2), e188-e194.

[4] Vo, A. H., Son, L. H., & Vo, M. T., Le, T. (2019). A novel framework for trash classification using deep transfer learning. *IEEE Access, 7,* 178631–178639.

[5] Sinha, B. B., & Dhanalakshmi, R. (2022). Recent advancements and challenges of Internet of Things in smart agriculture: A survey. *Future Generation Computer Systems*, *126*, 169-184.

[6] Aishwarya, M. P., & Reddy, A. P. (2024). Dataset of chilli and onion plant leaf images for classification and detection. *Data in Brief, 54,* 110524.

[7] Laizer, H., Mduma, N., Machuve, D., & Maganga, R. (2024). Common beans imagery dataset for early detection of bean rust and bean anthracnose diseases. *Data in Brief, 54,* 110508.

[8] Elfatimi, E., Eryigit, R., & Elfatimi, L. (2022). Beans leaf diseases classification using MobileNet models. *IEEE Access, 10,* 9471–9482.

[9] He, X., Fang, K., Qiao, B., Zhu, X., & Chen, Y. (2021). Watermelon disease detection based on deep learning. *International Journal of Pattern Recognition and Artificial Intelligence, 35(05),* 2152004.

[10] Alhazmi, S. (2023). Different stages of watermelon diseases detection using optimized CNN. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2022*, pp. 121-133.

[11] Santoso, H. A., Safsalta, B. F., Febrianto, N., Saraswati, G. W., & Haw, S. C. (2024). Comparative analysis of convolutional neural network and DenseNet121 transfer learning in agriculture focusing on crop leaf disease identification. *Applied Computing and Informatics*, (ahead-of-print), DOI: 10.1108/ACI-03-2024-0132

[12] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S. (2023). ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16133-16142.

[13] Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEiT: BERT Pre-Training of Image Transformers. *ICLR 2022*

[14] Shao, K., Wang, X., & Jiang, H. (2024). BEiT-based and feature extraction unsupervised clustering algorithm for wafer patterns. *Journal of Physics: Conference Series*, 2851(1), 012021.

[15] Ozdemir, B., & Pacal, I. (2025). An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms. *Results in Engineering*, 103692.

[16] Rajesh, N., Yalavarthy, A. B., Vamsi, V. K., & Saranya, G. (2024). BEiT Transformer Models to Aid in the Early Detection of Parkinson Illness. 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), pp. 1-7.
[17] Zhou, Z. H. (2025). Ensemble methods: foundations and algorithms. CRC press.
[18] Jiang, Y., Cao, Z., Wu, Y., Song, W., & Zhang, J. (2024). Ensemble-based deep reinforcement learning for vehicle routing problems under distribution shift. Advances in Neural Information Processing Systems, 36.
[19] Vasan, V., Sridharan, N. V., Balasundaram, R. J., & Vaithiyanathan, S. (2024). Ensemble-based deep learning model for welding defect detection and classification. Engineering Applications of Artificial Intelligence, 136, 108961.
[20] Nakib, M. I., & Mridha, M. F. (2024). Comprehensive watermelon disease recognition dataset. Data in Brief, 53, 110182.
[21] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, 4171-4186.
[22] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. ICML 2021, 8821-8831.

PHƯƠNG PHÁP HỌC SÂU DỰA TRÊN MÔ HÌNH TỔ HỢP SỬ DỤNG CONVNEXTV2 VÀ BEIT CHO PHÂN LOẠI BỆNH TRÊN DƯA HẤU

Lê Cung Tưởng

TÓM TẮT — Nghiên cứu này đề xuất MelonDx, một khung mô hình mới dựa trên phương pháp tổ hợp (ensemble) nhằm phân loại sớm các bệnh trên dưa hấu, hướng tới khắc phục những hạn chế và tính chủ quan của các phương pháp chẩn đoán truyền thống thông qua việc ứng dụng các mô hình học sâu tiên tiến. MelonDx kết hợp ConvNeXtV2 – một mạng nơ-ron tích chập (CNN) tiên tiến – với BEiT – một mô hình mã hóa transformer – thông qua kỹ thuật học tổ hợp để tận dụng thế mạnh bổ sung của cả hai trong trích xuất đặc trưng và phân loại bệnh. Khung mô hình này được đánh giá nghiêm ngặt trên bộ dữ liệu bệnh dưa hấu chuyên dụng, sử dụng các chỉ số như Accuracy, F1 score và AUC, và đạt hiệu suất vượt trội với độ chính xác 99,43%, F1 score 99,43% và AUC 99,99% - vượt qua cả các mô hình đơn lẻ và những phương pháp tiên tiến hiện có. Bằng cách kết hợp ưu thế của kiến trúc CNN và transformer, MelonDx nâng cao độ chính xác và độ tin cậy, mang đến một giải pháp tiềm năng cho nông nghiệp chính xác, giúp giảm thiểu thiệt hại mùa màng và tối ưu hóa hiệu quả sản xuất thông qua việc phát hiện bệnh kịp thời và chính xác.

Từ khóa — Học tổ hợp, Học sâu, Phân loại bệnh trên dưa hấu.

Tuong Le received his Ph.D. degree in Computer Science from Sejong University, Korea, in 2020. He is currently a researcher and faculty member at the Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Vietnam. He has authored over 40 publications in high-impact journals, including Information Sciences, Expert Systems with Applications, IEEE Access, and Engineering Applications of Artificial Intelligence. His research interests span machine learning, imbalanced learning, deep learning, business intelligence, data analysis, data mining, and pattern mining. He can be contacted at email: lc.tuong@hutech.edu.vn.