Black Sea Journal of Agriculture
doi: 10.47115/bsagriculture.1836017

Open Access Journal
e-ISSN: 2618 – 6578

Research Article
Volume 9 - Issue 3 : 332-344 / May 2026

CONVOLUTIONAL NEURAL NETWORKS FOR MULTI-CLASS WHEAT FUNGAL DISEASE DIAGNOSIS: A COMPARATIVE STUDY OF TRANSFER LEARNING MODELS

Cihat GEDİK[1]*, Bahadır SAYINCI[1], Taner YILDIZ[2]

[1]*Bilecik Seyh Edebali University, Faculty of Agriculture and Natural Sciences, Department of Biosystems Engineering, 11100, Bilecik, Türkiye*
[2]*Ondokuz Mayıs University, Faculty of Agriculture, Department of Agricultural Machinery and Technologies Engineering, 55139, Samsun, Türkiye*

Abstract: Wheat (*Triticum aestivum* L.) is a staple food for billions worldwide and plays a critical role in global food security. However, fungal pathogens cause yield losses and significant economic damage, posing a serious threat to wheat production. Therefore, early and accurate detection is strategically important to mitigate these losses. In this study, a dataset comprising 5,325 images of seven major wheat fungal diseases Blast, Brown Rust, Stripe Rust, Fusarium Head Blight, Loose Smut, Powdery Mildew, and Septoria along with healthy samples was used. The images were split into training (70%), validation (15%), and test (15%) subsets, and four pre-trained convolutional neural network (CNN) models, MobileNet, NASNetMobile, DenseNet121 and DenseNet169 were fine-tuned using transfer learning. Results showed that DenseNet169 achieved the highest classification accuracy (86.87%), followed by DenseNet121 (84.16%) and MobileNet (81.07%). NASNetMobile, however, demonstrated the lowest performance during validation. The findings highlight the strong potential of DenseNet169 in achieving high accuracy for wheat disease classification and its applicability in developing early detection systems that support sustainable agricultural production.

Keywords: Wheat diseases, Convolutional neural networks, Transfer learning, DenseNet169, Precision agriculture

***Corresponding author:** Bilecik Seyh Edebali University, Faculty of Agriculture and Natural Sciences, Department of Biosystems Engineering, 11100, Bilecik, Türkiye
E mail: cihat.gedik@bilecik.edu.tr (C. GEDİK)
Cihat GEDİK https://orcid.org/0000-0003-4955-2220
Bahadır SAYINCI https://orcid.org/0000-0001-7148-0855
Taner YILDIZ https://orcid.org/0000-0002-4774-6534

Received: December 04, 2025
Accepted: March 16, 2026
Published: May 15, 2026

Cite as: Gedik, C., Sayıncı, B., & Yıldız, T. (2026). Convolutional neural networks for multi-class wheat fungal disease diagnosis: A comparative study of transfer learning models. *Black Sea Journal of Agriculture*, 9(3): 332-344.

1. Introduction

Wheat (*Triticum aestivum* L.) is a vital staple crop that sustains billions of people and plays a central role in global food security. As a strategically important crop, its sustained production is critical for meeting rising food demand. However, fungal diseases continue to reduce yields and degrade grain quality, causing substantial economic losses (Figueroa et al., 2018; Langridge et al., 2022). Early and reliable diagnosis is therefore essential for timely interventions and sustainable agriculture (Duman, 2025).

Among the most common fungal diseases in wheat cultivation are Blast (*Magnaporthe oryzae*), Brown Rust (*Puccinia triticina*), Stripe Rust (*Puccinia striiformis*), Fusarium Head Blight (*Fusarium graminearum*), Loose Smut (*Ustilago tritici*), Powdery Mildew (*Blumeria graminis*), and Septoria (*Zymoseptoria tritici*). These pathogens cause significant yield and quality losses by adversely affecting physiological functions such as photosynthesis and grain development (Bai and Shaner, 2004; Bolton et al., 2008; Duba et al., 2018; Abrahim, 2019; Kumar and Kukreja, 2022; Martínez et al., 2022; Rana et al., 2022; Şahin et al., 2023). The selection of these seven fungal diseases was based on their widespread occurrence, economic impact, and diagnostic complexity.

Early disease diagnosis is crucial to prevent yield losses and support sustainable agriculture. In recent years, deep learning techniques have gained the capability to perform complex tasks in computer vision, such as object detection and image classification, with high accuracy, thus becoming an important tool for diagnosing agricultural diseases (Gerdan et al., 2023; Khalid and Karan, 2024). In addition to disease detection, image processing techniques are widely used in agricultural quality control processes, such as identifying structural defects like cracks and bruises in produce, as well as analyzing shape and size (Ercisli et al., 2012; Kara et al., 2013; Sayıncı et al., 2015). Convolutional Neural Networks (CNNs) are widely recognized as an effective tool for disease diagnosis by automatically detecting key features in images. Particularly in classification and recognition tasks, CNN models learn features automatically, providing high accuracy in computer

vision applications such as object detection and segmentation (LeCun et al., 1998; Krizhevsky et al., 2012; LeCun et al., 2015; Russakovsky et al., 2015; He et al., 2016).

Bouskour et al. (2024) applied data augmentation strategies based on the MobileNetV2 architecture to classify stripe rust and leaf blotch diseases in wheat, demonstrating that increasing the number of images significantly improved both training and validation performance. Long et al. (2023) developed the CerealConv model to classify four wheat diseases yellow rust, brown rust, powdery mildew, and Septoria along with healthy plants, achieving an accuracy rate of 97.05%. Reis and Türk (2023) combined an integrated deep learning framework with ensemble learning to classify yellow rust, brown rust, and healthy samples. Pan et al. (2022) developed a novel ensemble learning method named WR-EL for the rapid and accurate detection of stem rust and leaf rust. Shafi et al. (2023) integrated U2-Net-based leaf segmentation with the ResNet-50 model to detect stripe rust and classify infection severity, achieving 96% accuracy. Goyal et al. (2021) developed a model for classifying 10 different wheat diseases, achieving 97.88% test and 98.62% training accuracy. Mi et al. (2020) used the C-DenseNet model, enhanced with the Convolutional Block Attention Module (CBAM), to classify different infection levels of stripe rust, achieving 97.99% accuracy. Bao et al. (2021) reached 94.1% accuracy in a model integrating the CBAM module for detecting wheat ear diseases using a dataset consisting of two disease types and healthy samples. Nigam et al. (2024) achieved 98.7% test accuracy using the EfficientNet B0-CBAM model to detect fungal diseases such as stripe rust, leaf rust, and stem rust in the WheatRust21 image dataset.

Genaev et al. (2021) developed a system capable of classifying five different fungal wheat diseases using an EfficientNet-based model. Lu et al. (2017) proposed a mobile application capable of detecting six different diseases from the WDD2017 dataset, achieving 93.27% accuracy. Niaz et al. (2025) developed a mobile system capable of identifying 14 different wheat diseases based on machine learning algorithms. In light of the above studies, this research performs the classification of seven different wheat diseases and healthy wheat samples using MobileNet, NASNetMobile, DenseNet121, and DenseNet169 deep learning architectures based on transfer learning. The performances of different models are compared, and the advantages of transfer learning in disease diagnosis are revealed. This study demonstrates that deep learning-based approaches can enhance sustainable agricultural production by enabling accurate early diagnosis and classification of wheat diseases. Figure 1 illustrates the general architecture of a convolutional neural network, which forms the basis of the deep learning models evaluated in this study.

2. Materials and Methods

2.1. Image Dataset

The dataset used in this study was obtained from the publicly available 'Wheat Plant Diseases' dataset on Kaggle (https://www.kaggle.com/datasets/kushagra3204/wheat-plant-diseases). The Kaggle dataset was selected due to its public availability, standardized labeling, and frequent use in recent benchmarking studies. The original dataset comprises 15 classes representing various wheat diseases and pests. For the purposes of this study, only images belonging to the fungal disease classes illustrated in Figure 2 were selected for deep learning-based classification.

Based on preliminary experimental trials, the dataset was divided into training, validation, and test subsets to ensure reliable model development and evaluation, as summarized in Table 1. The dataset consists of a total of 5.325 images, with 3.195 images allocated to the training set, 1.064 to the validation set, and 1.066 to the test set, corresponding to 70%, 15%, and 15% of the dataset, respectively. The dataset includes seven fungal disease classes (Blast, Brown Rust, Fusarium Head Blight, Loose Smut, Powdery Mildew, Septoria, and Stripe Rust) along with healthy wheat samples.

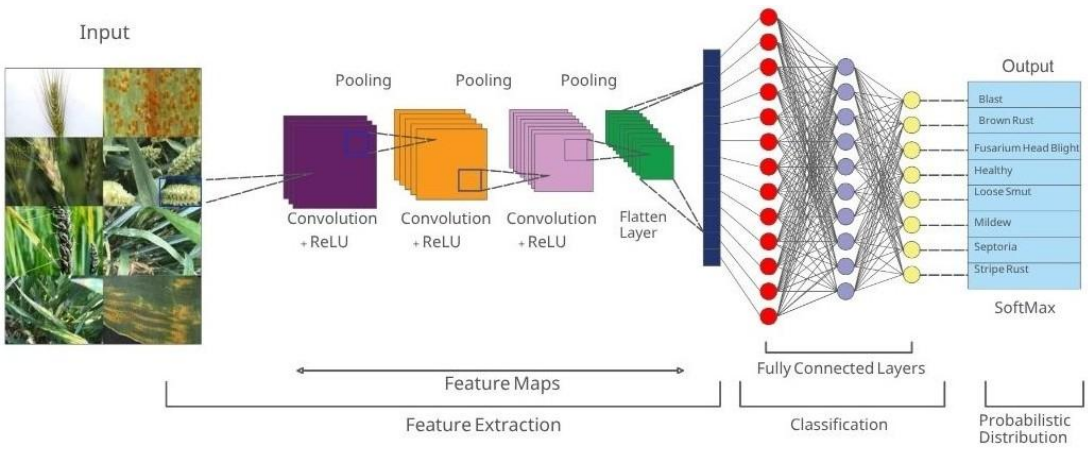

Figure 1. General architecture of a convolutional neural network (CNN) forming the basis of the models used for wheat disease classification.

Table 1. Distribution of images in the dataset across training, validation, and test subsets for each wheat disease class and healthy samples

Diseases	Train Set	Validation Set	Test Set	Total
Blast	221	74	74	369
Brown Rust	586	195	196	977
Fusarium Head Blight	360	119	120	599
Healthy	593	197	197	987
Loose Smut	249	83	83	415
Powdery Mildew	504	168	168	840
Septoria	207	69	69	345
Stripe Rust	475	158	159	792
Total	3195	1064	1066	5325

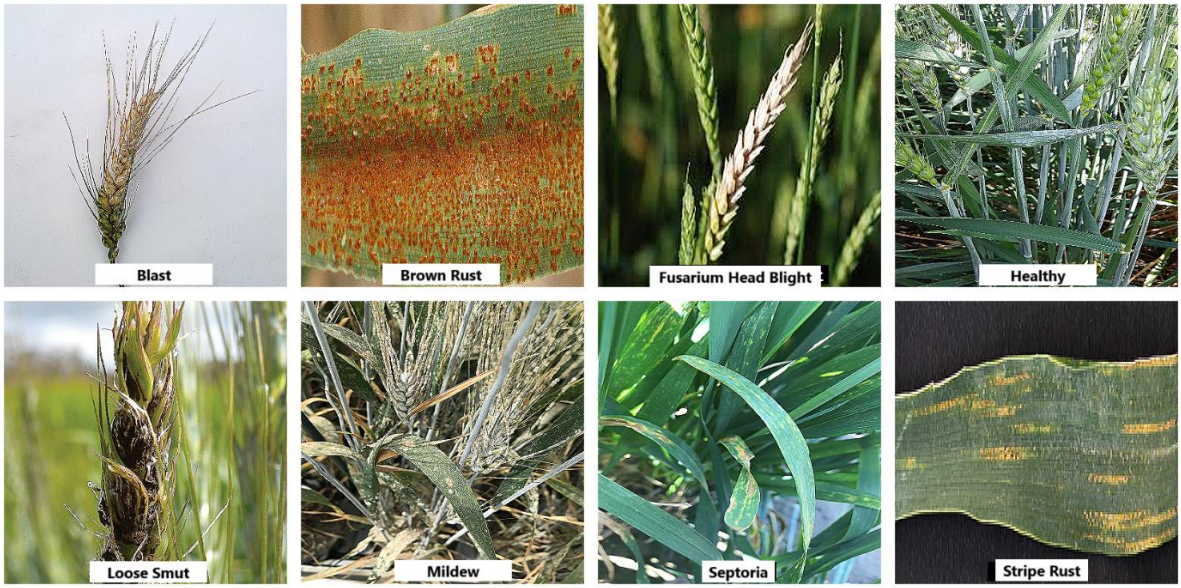

Figure 2. Representative images of the seven wheat fungal diseases (Blast, Brown Rust, Stripe Rust, Fusarium Head Blight, Loose Smut, Powdery Mildew and Septoria) and healthy wheat samples included in the dataset.

2.2. Data Augmentation

Data augmentation is widely used to enhance generalization and prevent overfitting in transfer learning-based deep learning models. Applying various transformations to the dataset increases both the number and diversity of training samples (Wang and Perez, 2017; Shorten and Khoshgoftaar, 2019). The parameter ranges used for augmentation in this study are summarized in Table 2. Specifically, random rotations within ±20° were applied to simulate natural camera angle variations. Horizontal and vertical shifts up to 20% of the image dimensions were used to improve spatial invariance. Both horizontal and vertical flipping were enabled to increase robustness against orientation differences. Zoom transformations ranging from 80% to 120% were applied to simulate variations in shooting distance. Additionally, shear transformations within ±20% were introduced to model slight perspective distortions. Reflective padding was used to fill newly created pixel regions after transformations to prevent artificial border artifacts.

Table 2. Data augmentation parameter settings

Method	Range
Rotation	±20°
Width shift	±%20
Height shift	±20%
Horizontal and Vertical flip	True
Zoom	80%–120%
Shear	±20%

2.3. Transfer Learning and Model Architectures

In this study, four pre-trained convolutional neural network architectures MobileNet, NASNetMobile, DenseNet121, and DenseNet169 were employed using a transfer learning approach. These models were selected to represent architectures with different depths, computational complexities, and parameter sizes, enabling a comparative evaluation of lightweight and deep network structures for wheat disease classification (Howard, 2017; Huang et al., 2017; Zoph et al., 2018). Pre-trained weights obtained from the ImageNet dataset were used for model initialization, and the final classification layers were replaced to accommodate the

eight target classes (Russakovsky et al., 2015; Shin et al., 2016).

During training, the convolutional base of each model was initially frozen, and only the newly added classification layers were trained. Subsequently, fine-tuning was performed by unfreezing the upper layers of the convolutional base to improve feature adaptation to the wheat disease dataset, following established transfer learning strategies (Pan and Yang, 2009; Weiss et al., 2016; Zhuang et al., 2020).

2.3.1. NASNetMobile

Developed by Zoph et al. (2018), NASNetMobile stands out for its automated neural architecture search capability. Designed to deliver high accuracy in image classification tasks and adaptability across datasets, the model leverages the Neural Architecture Search (NAS) principle to automate the architecture design process and generate more efficient deep learning models.

2.3.2. MobileNet

Introduced by Howard et al. (2017), MobileNet was one of the first deep learning architectures optimized for mobile devices with low computational requirements. Its core is built upon depthwise separable convolutions, which break the standard convolution process into two steps: depthwise convolution (applying a filter to each input channel separately) and pointwise convolution (1×1 convolution combining the results). This structure reduces computation while maintaining competitive accuracy, making it suitable for resource-constrained environments.

2.3.3. DenseNet

Proposed by Huang et al. (2017), the DenseNet architecture establishes dense connections between all layers, enabling each layer to receive direct information from all preceding layers. This design mitigates the vanishing gradient problem, strengthens feature propagation, and accelerates learning. DenseNet variants such as DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264 differ in network depth, with deeper variants generally providing higher representational capacity (Pan et al., 2019).

2.4. Hyperparameters and Fine-Tuning Strategies

In addition to the baseline hyperparameters presented in Table 3, several optimization strategies were incorporated to enhance model generalization and convergence stability. A dynamic learning rate strategy was employed. Specifically, an exponential decay scheduler reduced the learning rate by 4% at the end of each epoch. In addition, ReduceLROnPlateau was applied to decrease the learning rate by a factor of 0.1 when validation loss plateaued, with a minimum threshold of 1e-7. This combined scheduling strategy facilitated smoother convergence during extended training.

To mitigate potential class imbalance, class weights were computed using a balanced weighting scheme and integrated into the categorical cross-entropy loss function. This adjustment reduced bias toward dominant classes and improved minority class sensitivity.

Training was conducted in two stages. In the initial transfer learning phase, the convolutional layers of each pre-trained backbone architecture were frozen to preserve previously learned generic feature representations, while only the newly added fully connected classification layers were trained. In the subsequent fine-tuning phase, all layers were unfrozen and jointly optimized using a reduced learning rate of 1e-5. Training was performed for up to 300 epochs with early stopping (patience = 20) based on validation loss.

Table 3. Hyperparameters and training settings used for NASNetMobile, MobileNet, DenseNet121, and DenseNet169 models in wheat disease classification.

Hyperparameter	Value / Settings	Description
Input Shape	(224, 224, 3)	Standard RGB image size
Number of Classes	8	Multi-class wheat disease classification
Batch Size	64	Determined based on GPU memory capacity
Optimizer	AdamW	Adaptive optimization with L2 regularization
Weight Decay	1.00E-04	L2 regularization strength
Initial Learning Rate	1.00E-04	Transfer learning phase
Fine-Tuning Learning Rate	1.00E-05	Full network optimization phase
Learning Rate Scheduler	Exponential decay (×0.96 per epoch)	Gradual reduction of learning rate
ReduceLROnPlateau	Factor=0.1, Patience=20, Min_lr=1e-7	Dynamic LR reduction based on validation loss
Epochs	Up to 300	Maximum training duration
Early Stopping	Patience=20, Monitor=val_loss	Prevents overfitting
Loss Function	Categorical Crossentropy	Multi-class classification objective
Class Weights	Balanced weighting scheme	Handles class imbalance
Dense Layers	256 + 128 neurons (ReLU)	Enhanced classification head
Batch Normalization	After each Dense layer	Improves training stability
Dropout	0.5 (first FC), 0.3 (second FC)	Reduces overfitting
Rescaling	1/255 normalization	Pixel value normalization
Base Model Strategy	Freeze → Unfreeze	Two-phase transfer learning approach

The AdamW optimizer with a weight decay of 1e-4 was used to implement L2 regularization.

To improve performance, Early Stopping was implemented with a patience of 20 epochs based on validation loss, and ReduceLROnPlateau was used to reduce the learning rate by a factor of 0.1 when validation loss plateaued, with a minimum learning rate of 1e-7. The output layer used GlobalAveragePooling2D, followed by a dense layer with 256 neurons and ReLU activation, and a 50% dropout rate to prevent overfitting. Model performance was evaluated using accuracy as the primary metric.

2.5. Performance Metrics

Performance metrics are essential for evaluating the classification capability of deep learning models. In this study, multiple complementary evaluation metrics were employed to ensure a comprehensive assessment of model performance.

Precision is defined as the proportion of correctly predicted positive samples among all predicted positive samples (Davis and Goadrich, 2006), as given in Equation 1:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall (Sensitivity) represents the proportion of correctly identified positive samples among all actual positive samples (Davis and Goadrich, 2006), as shown in Equation 2:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The F1-score, defined as the harmonic mean of precision and recall (Goutte and Gaussier, 2005), is calculated according to Equation 3:

$$F1 = 2x\frac{(Precision \cdot Recall)}{(Precision + Recall)} \quad (3)$$

Overall classification accuracy is computed as indicated in Equation 4:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

For multi-class classification, class-wise metrics were calculated using a one-vs-rest strategy, and the macro F1-score was obtained as the unweighted mean of individual class F1-scores (Too et al., 2019).

To address potential class imbalance, Balanced Accuracy was calculated as the mean recall across all classes, as defined in Equation 5:

$$Balanced\ Accuracy = \frac{1}{C}\sum_{i=1}^{C} Recall_i \quad (5)$$

During training, the categorical cross-entropy loss (LCCE), defined in Equation 6, was used to quantify the discrepancy between predicted probabilities and ground-truth labels (Goutte and Gaussier, 2005):

$$LCCE = -\sum_{i=1}^{C} y_i \log(p_i) \quad (6)$$

The Receiver Operating Characteristic (ROC) curve is a graphical tool used to evaluate the discriminative ability of a classification model across different decision thresholds. It illustrates the trade-off between sensitivity (True Positive Rate, TPR) and specificity by plotting TPR against the False Positive Rate (FPR). This representation enables threshold-independent assessment of model performance. The Area Under the ROC Curve (AUC) provides a single scalar metric summarizing the classifier's overall discriminative capability, independent of any specific threshold. An AUC value close to 1 indicates excellent separability between classes, whereas a value near 0.5 reflects performance comparable to random guessing (Park et al., 2004; Hoo et al., 2017). The TPR and FPR are formally defined in Equations 7 and 8, respectively.

$$TPR = \frac{TP}{TP + FN} \quad (7)$$

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

The Area Under the ROC Curve (AUC), given in Equation 9, summarizes the model's discriminative capability:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \quad (9)$$

For multi-class classification, ROC curves were computed using the one-vs-rest approach, and macro-average AUC values were reported.

Where, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, y_i: ground-truth label, p_i: predicted probability, C: number of classes.

Model predictions were further analyzed using a confusion matrix. For each class, a one-vs-rest (OvR) strategy was adopted, where the selected class was treated as the positive class and all remaining classes were grouped as the negative class (Sathyanarayanan and Tantri, 2024). Based on this formulation, True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) were computed for each class individually, as illustrated in Figure 3.

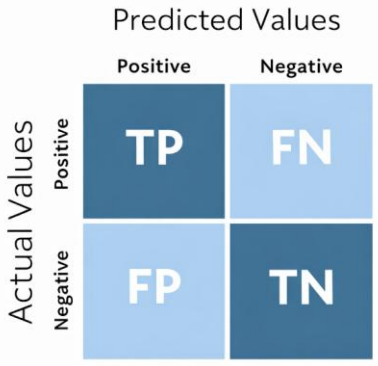

Figure 3. Structure of a binary confusion matrix used to derive classification performance metrics.

2.6. Experimental Setup

The model training was conducted in a cloud-based GPU environment equipped with an NVIDIA A100 graphics processor, CUDA 11.4, and TensorFlow 2.4.1.

3. Results

3.1. Performance of CNN Models

The performance of the evaluated CNN architectures was analyzed in terms of both training behavior and final classification results. Figure 4 presents the training and validation accuracy and loss curves of NASNetMobile, MobileNet, DenseNet121, and DenseNet169 models, providing insight into the convergence characteristics and optimization stability during the learning process.

Final model performance was evaluated exclusively on the independent test dataset to ensure an unbiased assessment of generalization capability. As summarized in Table 4, DenseNet169 achieved the highest test accuracy (86.87%), followed by DenseNet121 (84.16%) and MobileNet (81.07%), while NASNetMobile demonstrated the lowest test accuracy (78.99%). In addition to overall accuracy, DenseNet169 also yielded the highest balanced accuracy (87.06%) and macro F1-score (85.57%), indicating superior performance across both majority and minority classes.

Considering test accuracy, loss values, and training duration, DenseNet169 emerged as the most effective architecture for wheat disease classification. DenseNet121 also demonstrated strong performance, whereas MobileNet provided a favorable balance between computational efficiency and classification accuracy. In contrast, NASNetMobile showed comparatively weaker performance under the same experimental conditions.

3.2. Performance Metrics Results

The class-wise Precision, Recall, and F1-score values obtained from the four transfer learning architectures are presented in Table 5, while their graphical comparison is illustrated in Figure 5.

As shown in both Table 5 and Figure 4, Brown Rust was the most consistently classified disease across all models. F1-scores exceeded 92% for every architecture and reached 98.19% with DenseNet169, indicating strong separability and stable detection performance. Similarly, Fusarium Head Blight and Loose Smut demonstrated balanced Precision and Recall values, particularly under the DenseNet-based models. In contrast, Septoria exhibited the greatest performance imbalance. Although Recall values were relatively high (78.57%–94.29%), Precision remained comparatively low (44.00%–56.48%) across models.

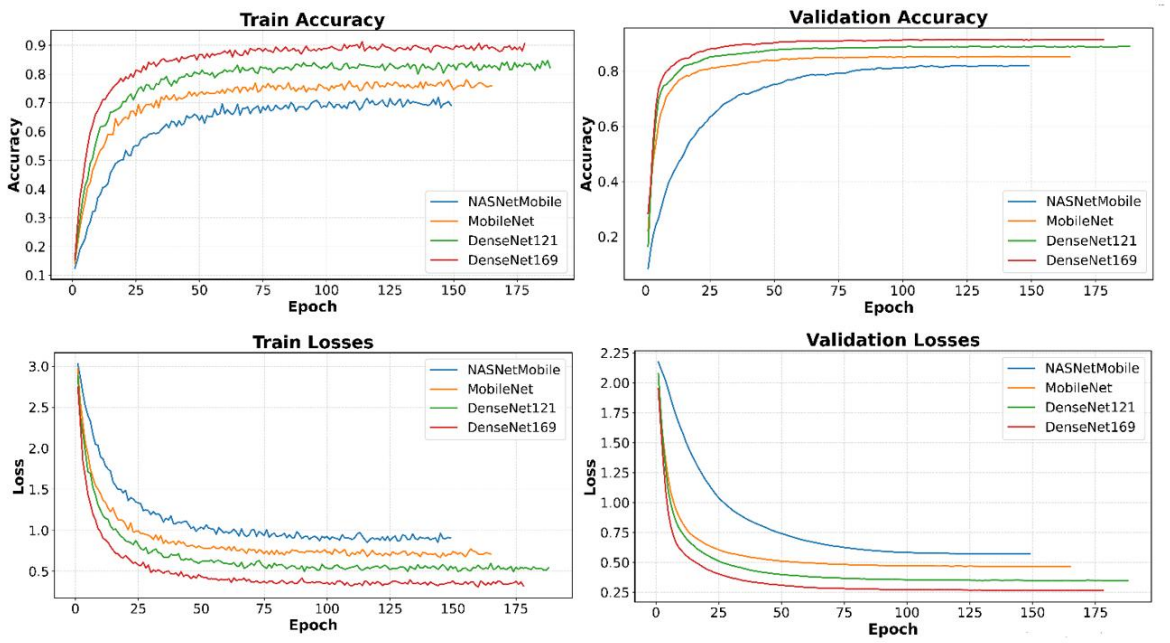

Figure 4. Training, validation, and test accuracy and loss curves for NASNetMobile, MobileNet, DenseNet121, and DenseNet169 models.

Table 4. Test performance results of the evaluated CNN models

Model	Accuracy	Balanced Accuracy	Macro F1	Loss	Time (min)
NASNetMobile	0.7899	0.7921	0.7730	0.5545	294
MobileNet	0.8107	0.8165	0.7947	0.4340	260
DenseNet121	0.8416	0.8497	0.8266	0.3616	262
DenseNet169	0.8687	0.8706	0.8557	0.3057	224

Table 5. Precision, Recall, and F1-Score metrics across CNN models in wheat disease classification

Disease	Metric	NASNetMobile	MobileNet	DenseNet121	DenseNet169
Blast	Precision	0.5481	0.5752	0.6078	0.7674
	Recall	0.7703	0.8784	0.8378	0.8919
	F1-Score	0.6404	0.6952	0.7045	0.825
Brown Rust	Precision	0.9773	0.989	0.9841	0.9948
	Recall	0.8776	0.9184	0.949	0.9694
	F1-Score	0.9247	0.9524	0.9662	0.9819
Fusarium Head Blight	Precision	0.8824	0.9469	0.906	0.9478
	Recall	0.875	0.8917	0.8833	0.9083
	F1-Score	0.8787	0.9185	0.8945	0.9277
Healthy	Precision	0.7619	0.8	0.8113	0.7949
	Recall	0.8122	0.8122	0.8731	0.9442
	F1-Score	0.7862	0.806	0.8411	0.8631
Loose Smut	Precision	0.8025	0.8353	0.8706	0.8824
	Recall	0.7831	0.8554	0.8916	0.9036
	F1-Score	0.7927	0.8452	0.881	0.8929
Powdery Mildew	Precision	0.8321	0.9407	0.9508	0.9669
	Recall	0.6488	0.6607	0.6905	0.6964
	F1-Score	0.7291	0.7762	0.8	0.8097
Septoria	Precision	0.4957	0.44	0.5455	0.5648
	Recall	0.8406	0.7857	0.9429	0.8841
	F1-Score	0.6237	0.5641	0.6911	0.6893
Stripe Rust	Precision	0.9062	0.8855	0.9748	0.9683
	Recall	0.7296	0.7296	0.7296	0.7673
	F1-Score	0.8084	0.8	0.8345	0.8561

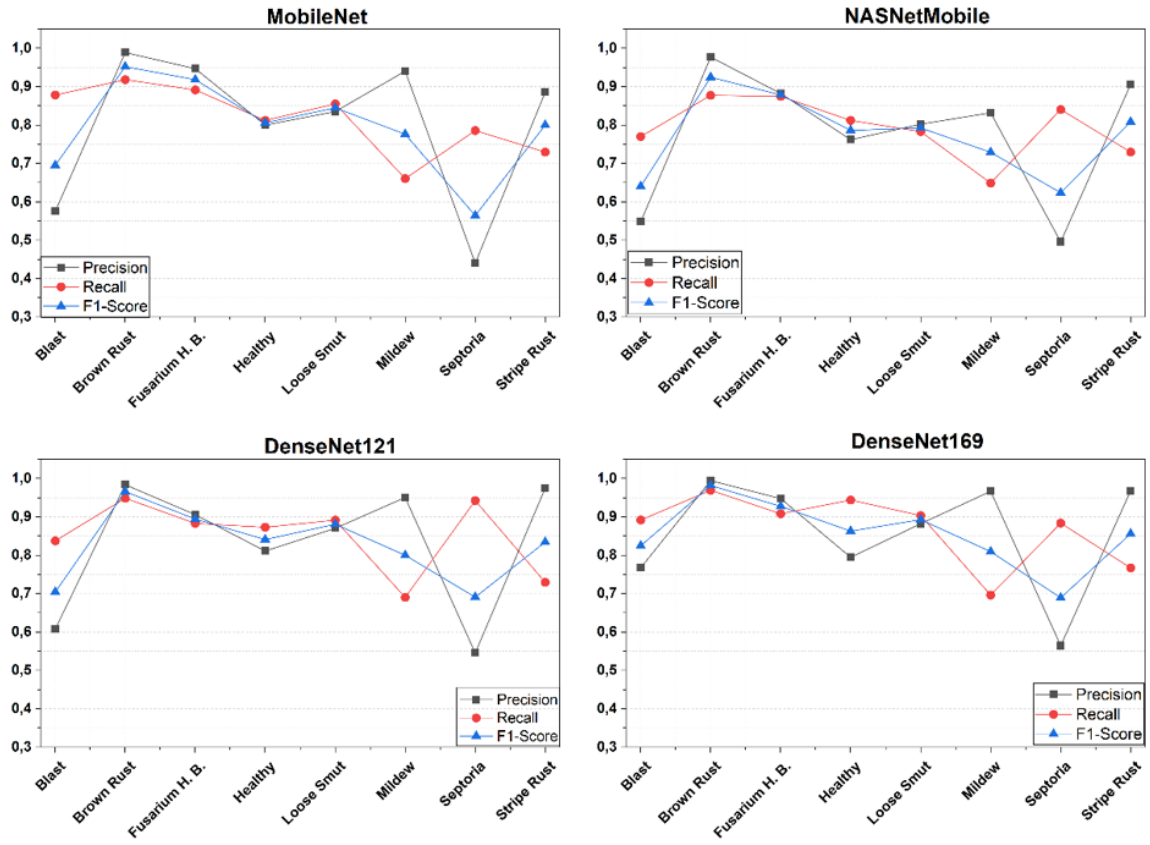

Figure 5. Precision, Recall, and F1-Score results for wheat disease classification across four CNN models.

This disparity, clearly visible in Figure 5, suggests a tendency toward false positive predictions and highlights the classification difficulty of this class. A comparable but less pronounced imbalance was observed for Powdery Mildew, where high Precision values (up to 96.69%) were accompanied by moderate Recall (64.88%–69.64%), indicating missed detections.

Among the evaluated architectures, DenseNet169 achieved the most consistent overall performance. As illustrated in Figure 4, it produced the highest F1-scores in five of the eight disease categories (Brown Rust, Fusarium Head Blight, Loose Smut, Healthy, and Stripe Rust). DenseNet121 also showed stable behavior across most classes but remained slightly below DenseNet169 in overall consistency. In comparison, NASNetMobile and MobileNet displayed greater variability in challenging classes such as Septoria and Powdery Mildew.

Overall, both Table 5 and Figure 5 confirm that DenseNet-based architectures provide improved class-wise stability and stronger F1-score consistency in multi-class wheat disease classification, with DenseNet169 emerging as the most effective model among the evaluated approaches.

3.3. Confusion Matrix Results

To further evaluate classification performance, confusion matrices for each model based on the test dataset are presented in Figure 6. In these matrices, the vertical axis represents the true category and the horizontal axis represents the predicted category. Class indices are defined as follows: 0– Blast, 1– Brown Rust, 2– Fusarium Head Blight, 3– Healthy, 4– Loose Smut, 5– Powdery Mildew, 6– Septoria, and 7– Stripe Rust.

Across all models, Brown Rust exhibited the most consistent and accurate classification performance, with particularly high true positive counts in DenseNet169 (193) and DenseNet121 (186). Healthy and Fusarium Head Blight were also classified with high reliability, especially in DenseNet-based architectures, indicating strong feature discrimination capability.

Despite the relatively limited number of Septoria samples, all models achieved acceptable true positive predictions for this class. Overall, Brown Rust, Fusarium Head Blight, and Loose Smut demonstrated strong class separability, while Healthy and Stripe Rust maintained stable performance with minor cross-class confusion. In contrast, Powdery Mildew and Septoria showed moderate inter-class misclassification, reflecting higher classification difficulty compared to other disease categories.

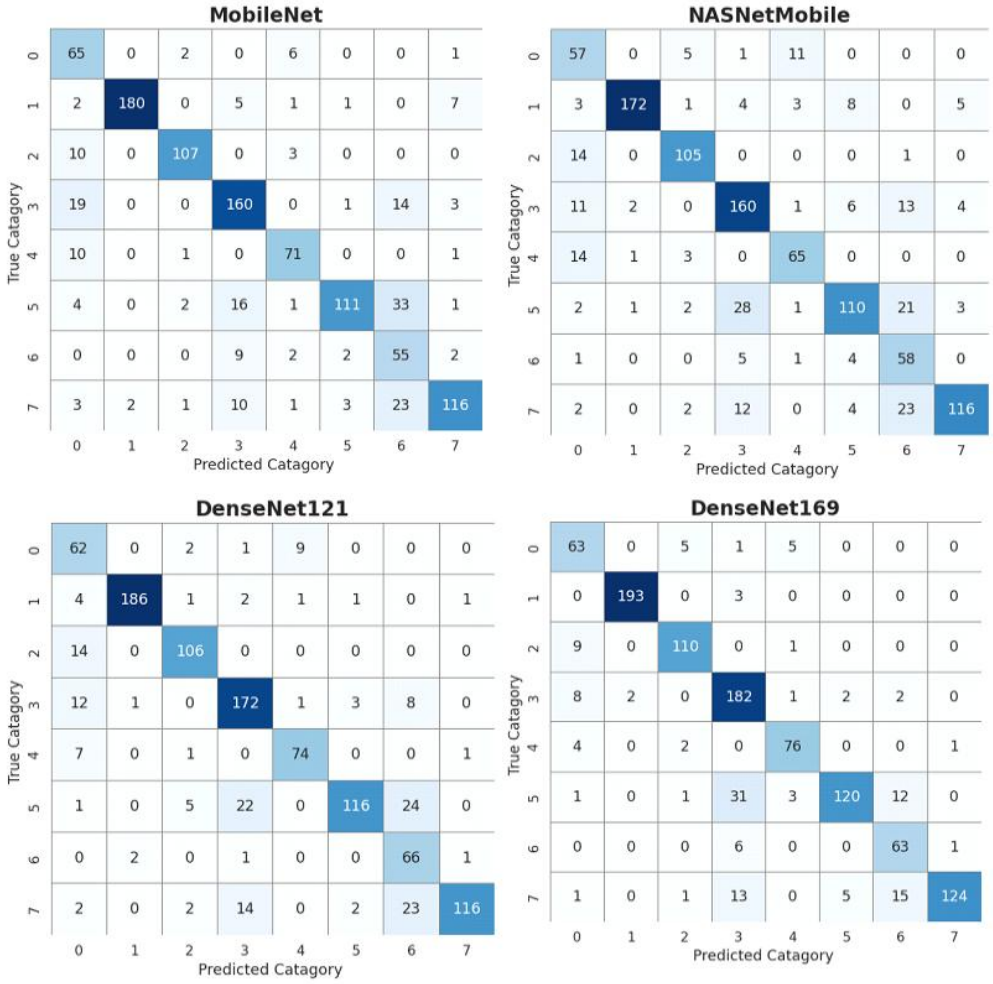

Figure 6. Confusion matrices of NASNetMobile, MobileNet, DenseNet121, and DenseNet169 models on the test dataset for eight wheat classes (seven fungal diseases and healthy samples).

3.4. ROC Curve Analysis

The ROC curves presented in Figure 7 demonstrate the class-wise discriminative performance of the evaluated CNN architectures on the test dataset. Across all models, the curves are positioned well above the diagonal reference line, indicating strong classification capability compared to random guessing.

DenseNet169 achieved the highest and most consistent AUC values across nearly all classes, with Brown Rust reaching an AUC of 1.000 and Septoria achieving 0.990. Similarly, Fusarium Head Blight (0.993), Loose Smut (0.992), and Stripe Rust (0.988) demonstrated excellent separability. These results confirm the superior discriminative stability of DenseNet169. DenseNet121 also showed strong performance, with AUC values exceeding 0.98 for most classes and reaching 0.999 for Brown Rust.

NASNetMobile and MobileNet achieved high AUC scores for dominant classes such as Brown Rust and Fusarium Head Blight; however, relatively lower AUC values were observed for certain classes, particularly Stripe Rust and Septoria, indicating comparatively reduced separability. Overall, all models achieved AUC values substantially above 0.90 for the majority of disease categories, demonstrating strong multi-class discrimination capability.

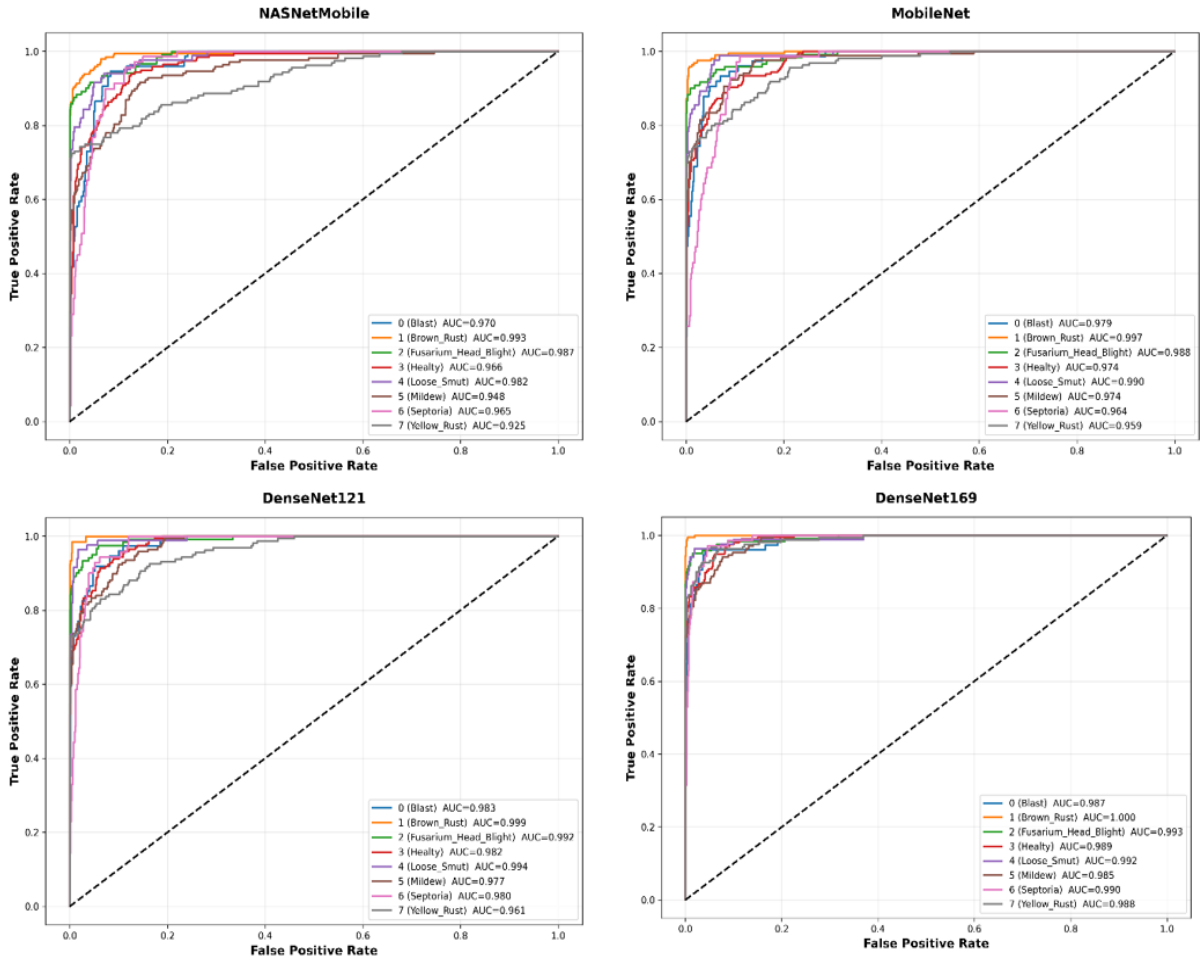

Figure 7. ROC curve analysis results.

3.5. Class-wise Grad-CAM Analysis

Grad-CAM was applied to representative correctly classified samples from each class to assess whether the network attends to disease-specific morphological features. The activation maps indicate that DenseNet169 predominantly focuses on symptomatic regions such as rust pustules, spike bleaching, necrotic lesions, and fungal accumulations, depending on the disease category. For classes with high classification performance (e.g., Brown Rust and Fusarium Head Blight), the activations are sharply localized around infected tissues, indicating strong discriminative feature learning (Figure 8). In contrast, comparatively challenging classes such as Septoria and Powdery Mildew exhibit more spatially diffuse activation patterns, suggesting partial overlap of visual features across categories, which aligns with the moderate inter-class confusion observed in the confusion matrix results.

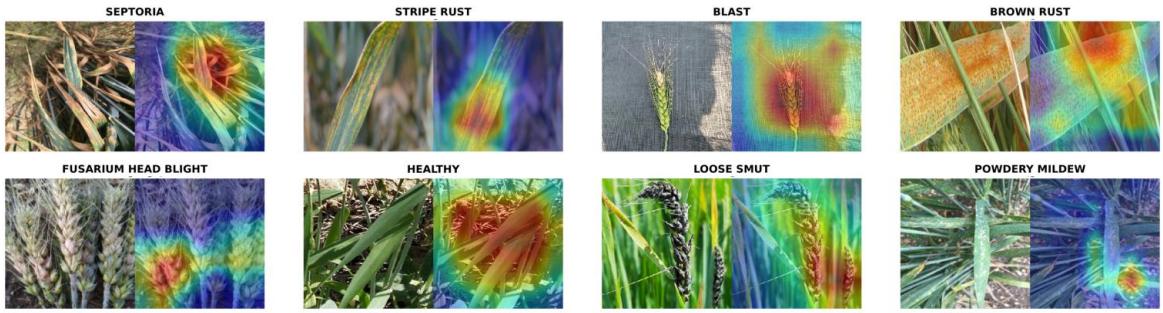

Figure 8. Grad-CAM activation maps generated by DenseNet169 for correctly classified samples of eight wheat disease classes.

4. Discussion

This study evaluated the classification performance of various CNN architectures for wheat fungal disease detection, and the obtained accuracy rates were compared with those reported in similar studies in the literature (Table 6). DenseNet121 achieved an accuracy of 84.16%, which is slightly lower than the 90.36% reported by Chang et al. (2024). This difference can be attributed to variations in dataset composition, image quality, class diversity, and preprocessing procedures. Notably, the present study involves eight classes (seven fungal diseases and healthy samples), increasing classification complexity compared to studies with fewer categories.

MobileNet achieved an accuracy of 81.07%, which is lower than the performance reported by Bouskour et al. (2024) and Jiang et al. (2022) using MobileNetV2-based architectures. These discrepancies are likely related to differences in the number of disease classes, disease types considered, dataset size, and data augmentation strategies, all of which substantially influence model generalization performance. The highest classification performance was obtained with DenseNet169, achieving a test accuracy of 86.87%, outperforming all other evaluated architectures. In contrast, Pan et al. (2022) reported an accuracy of 81% using the same model. The superior performance of DenseNet169 may be attributed to its deeper architecture and dense connectivity mechanism, which facilitates effective feature reuse and improved gradient propagation (Huang et al., 2017). By enabling each layer to access feature maps from all preceding layers, DenseNet169 enhances hierarchical feature representation and mitigates the vanishing gradient problem. This structural advantage is particularly beneficial in multi-class plant disease classification tasks, where subtle inter-class visual similarities require fine-grained discriminative learning.

The ROC curve analysis further confirmed the discriminative robustness of DenseNet-based architectures. DenseNet169 demonstrated consistently high AUC values across nearly all classes, indicating strong threshold-independent classification capability. This suggests that the model maintains high sensitivity and specificity across varying decision thresholds. The combination of high balanced accuracy and macro F1-score further indicates that DenseNet169 preserves stable performance not only for dominant classes but also for less represented categories.

Table 6. Comparative analysis of recently published studies

Reference	Models	Diseases	Number of Images	Accuracy Rates (%)
Chang et al. (2024)	DenseNet121	Stripe Rust, Leaf Rust, Stem Rust and Healthy	5251	90.36
Pan et al. (2022)	DenseNet169	Stem rust, Leaf rust and Healthy	810	81
Bouskour et al. (2024)	MobileNetV2	Septoria leaf, Stripe rust leaf and Healthy	1345	99
Jiang et al. (2022)	DenseNet121 MobileNetV3	Powdery mildew, Leaf rust, Stripe rust and Healthy	2643	92 88.08
Proposed	DenseNet169	Blast, Brown rust, Fusarium head blight, Healthy, Loose smut, Powdery mildew, Septoria and Stripe rust	5325	86.87

The comparatively lower classification performance observed for the Septoria class is likely influenced by its limited representation in the dataset. Class imbalance is a well-known challenge in multi-class classification tasks, as models tend to be biased toward classes with higher sample frequencies. Similar observations have been reported in previous plant disease classification studies, where underrepresented disease classes exhibited reduced recall and increased misclassification rates (Mohanty et al., 2016; Too et al., 2019). Although class

weighting strategies were applied in this study, fully eliminating imbalance effects remains challenging, particularly when visual symptoms overlap across disease categories. The dataset used in this study was obtained from a publicly available Kaggle repository, which may introduce certain limitations. Although standardized labeling is provided, the dataset exhibits class imbalance, with certain disease categories being underrepresented. In addition, variability in image acquisition conditions such as differences in illumination, background complexity, camera angles, and overall image quality may negatively affect model generalization. These uncontrolled factors can introduce noise and bias into the learning process, potentially limiting the robustness of the trained models under real-field conditions.

5. Conclusion

This study evaluated four transfer learning-based CNN architectures (NASNetMobile, MobileNet, DenseNet121, and DenseNet169) for multi-class wheat fungal disease classification. Among the evaluated models, DenseNet169 achieved the highest test accuracy and demonstrated stable performance across precision, recall, and F1-score metrics, particularly for Brown Rust and Fusarium Head Blight.

Comparative analysis with existing studies indicated that the obtained results are competitive with, and in some cases superior to, previously reported performances. DenseNet-based architectures showed consistent classification capability, while MobileNet provided a favorable trade-off between accuracy and computational efficiency, suggesting suitability for resource-constrained deployment.

Despite high classification accuracy for several disease classes, lower performance was observed for Powdery Mildew, and occasional confusion between Septoria and Powdery Mildew was identified. These findings indicate that further improvements may be achieved through enhanced dataset balance and class representation.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	C.G.	B.S.	T.Y.
C	40	30	30
D	40	30	30
S	-	50	50
DCP	80	10	10
DAI	25	50	25
L	40	20	40
W	40	20	40
CR	-	50	50
SR	50	-	50
PM	10	80	10

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

Abrahim, A. (2019). Loose smut of wheat (Ustilago tritici) and its managements. *Journal of Biology, Agriculture and Healthcare, 9*(8), 25-33.

Bai, G., & Shaner, G. (2004). Management and resistance in wheat and barley to Fusarium head blight. *Annual Review of Phytopathology, 42*(1), 135–161. https://doi.org/10.1146/annurev.phyto.42.040803.140340

Bao, W., Yang, X., Liang, D., Hu, G., & Yang, X. (2021). Lightweight convolutional neural network model for field wheat ear disease identification. *Computers and Electronics in Agriculture, 189*, 106367. https://doi.org/10.1016/j.compag.2021.106367

Bolton, M. D., Kolmer, J. A., & Garvin, D. F. (2008). Wheat leaf rust caused by Puccinia triticina. *Molecular Plant Pathology, 9*(5), 563–575. https://doi.org/10.1111/j.1364-3703.2008.00487.x

Bouskour, S., Zaggaf, M. H., & Bahatti, L. (2024). Deep learning recognition of wheat leaf disease using MobileNetV2 model. In *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (pp. 1–6). IEEE. https://doi.org/10.1109/IRASET60544.2024.10548207

Chang, S., Yang, G., Cheng, J., Feng, Z., Fan, Z., Ma, X., & Zhao, C. (2024). Recognition of wheat rusts in a field environment based on improved DenseNet. *Biosystems Engineering*, 238, 10-21. https://doi.org/10.1016/j.biosystemseng.2023.12.016

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *the 23rd International Conference on Machine Learning* (pp. 233–240). https://doi.org/10.1145/1143844.1143874

Duba, A., Goriewa-Duba, K., & Wachowska, U. (2018). A review of the interactions between wheat and wheat pathogens: Zymoseptoria tritici, Fusarium spp. and Parastagonospora

nodorum. *International Journal of Molecular Sciences, 19*(4), 1138. https://doi.org/10.3390/ijms19041138

Duman, B. (2025). Mobile device-based detection system of diseases and pests in rose plants using deep convolutional neural networks and quantization. *Journal of Agricultural Sciences, 31*(2), 302–318. https://doi.org/10.15832/ankutbd.1514972

Ercisli, S., Sayinci, B., Kara, M., Yildiz, C., & Ozturk, I. (2012). Determination of size and shape features of walnut (Juglans regia L.) cultivars using image processing. *Scientia Horticulturae, 133*, 47–55. https://doi.org/10.1016/j.scienta.2011.10.014

Figueroa, M., Hammond-Kosack, K. E., & Solomon, P. S. (2018). A review of wheat diseases: A field perspective. *Molecular Plant Pathology, 19*(6), 1523–1536. https://doi.org/10.1111/mpp.12618

Genaev, M. A., Skolotneva, E. S., Gultyaeva, E. I., Orlova, E. A., Bechtold, N. P., & Afonnikov, D. A. (2021). Image-Based Wheat Fungi Diseases Identification by Deep Learning. Plants, 10(8), 1500. https://doi.org/10.3390/plants10081500

Gerdan, D., Koç, C., & Vatandaş, M. (2023). Diagnosis of tomato plant diseases using pre-trained architectures and a proposed convolutional neural network model. *Journal of Agricultural Sciences, 29*(2), 618–629. https://doi.org/10.15832/ankutbd.957265

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval* (pp. 345–359). Springer Berlin Heidelberg.

Goyal, L., Sharma, C. M., Singh, A., & Singh, P. K. (2021). Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture. *Informatics in Medicine Unlocked*, 25, 100642. https://doi.org/10.1016/j.imu.2021.100642

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve?. *Emergency Medicine Journal, 34*(6), 357-359.

Howard, A. G. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv Preprint, arXiv:1704.04861*. https://doi.org/10.48550/arXiv.1704.04861

Huang, G., Liu, Z., Van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708).

Jiang J, Liu H, Zhao C, He C, Ma J, Cheng T, Zhu Y, Cao W, Yao X. Evaluation of Diverse Convolutional Neural Networks and Training Strategies for Wheat Leaf Disease Identification with Field-Acquired Photographs. *Remote Sensing*. 2022; 14(14):3446. https://doi.org/10.3390/rs14143446

Kara, M., Sayıncı, B., Elkoca, E., Öztürk, İ., & Özmen, T. (2013). Seed size and shape analysis of registered common bean (Phaseolus vulgaris L.) cultivars in Turkey using digital photography. *Journal of Agricultural Sciences*, 19(3), 219–234. https://doi.org/10.1501/Tarimbil_0000001247

Khalid, M. M., & Karan, O. (2024). Deep learning for plant disease detection. *International Journal of Mathematics, Statistics, and Computer Science, 2*, 75–84. https://doi.org/10.59543/ijmscs.v2i.8343

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*. https://doi.org/10.1145/3065386

Kumar, D., & Kukreja, V. (2022). Deep learning in wheat diseases classification: A systematic review. *Multimedia Tools and Applications, 81*(7), 10143–10187. https://doi.org/10.1007/s11042-022-12160-3

Langridge P, Alaux M, Almeida NF, Ammar K, Baum M, Bekkaoui F, Bentley AR, Beres BL, Berger B, Braun H-J, et al. Meeting the Challenges Facing Wheat Production: The Strategic Research Agenda of the Global Wheat Initiative. *Agronomy*. 2022; 12(11):2767. https://doi.org/10.3390/agronomy12112767

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Long, M., Hartley, M., Morris, R. J., & Brown, J. K. (2023). Classification of wheat diseases using deep learning networks with field and glasshouse images. *Plant Pathology*, 72(3), 536-547. https://doi.org/10.1111/ppa.13684

Lu, J., Hu, J., Zhao, G., Mei, F., & Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. Computers and *Electronics in Agriculture, 142*, 369–379. https://doi.org/10.1016/j.compag.2017.09.012

Martínez, M., Biganzoli, F., Arata, A., Dinolfo, M. I., Rojas, D., Cristos, D., & Stenglein, S. (2022). Warm nights increase Fusarium head blight negative impact on barley and wheat grains. *Agricultural and Forest Meteorology, 318*, 108909. https://doi.org/10.1016/j.agrformet.2022.108909

Mi, Z., Zhang, X., Su, J., Han, D., & Su, B. (2020). Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. *Frontiers in Plant Science, 11*, 558126. https://doi.org/10.3389/fpls.2020.558126

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*, 1419. https://doi.org/10.3389/fpls.2016.01419

Niaz, A. A., Ashraf, R., Mahmood, T., Faisal, C. N., & Abid, M. M. (2025). An efficient smartphone application for wheat crop diseases detection using advanced machine learning. *PLOS One, 20*(1), e0312768. https://doi.org/10.1371/journal.pone.0312768

Nigam, S., Jain, R., Singh, V. K., Marwaha, S., Arora, A., & Jain, S. (2024). EfficientNet architecture and attention mechanism-based wheat disease identification model. *Procedia Computer Science, 235*, 383–393. https://doi.org/10.1016/j.procs.2024.04.038

Pan, Q., Gao, M., Wu, P., Yan, J., & AbdelRahman, M. A. (2022). Image classification of wheat rust based on ensemble learning. *Sensors, 22*(16), 6047. https://doi.org/10.3390/s22166047

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pan, W., Qin, J., Xiang, X., Wu, Y., Tan, Y., & Xiang, L. (2019). A smart mobile diagnosis system for citrus diseases based on densely connected convolutional networks. *IEEE Access, 7*, 87534–87542. https://doi.org/10.1109/ACCESS.2019.2924973

Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean journal of radiology, 5*(1), 11-18.

Rana, V., Batheja, A., Sharma, R., Rana, A., & Priyanka. (2022). Powdery mildew of wheat: Research progress, opportunities, and challenges. In *New Horizons in Wheat and Barley Research: Crop Protection and Resource Management* (pp. 133–178).

Reis, H. C., & Türk, V. (2024). Integrated deep learning and ensemble learning model for deep feature-based wheat disease detection. *Microchemical Journal*, 197, 109790. https://doi.org/10.1016/j.microc.2023.109790

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Şahin, Y., Bütüner, A., & Erdoğan, H. (2023). Potential for early detection of powdery mildew in okra under field conditions using thermal imaging. *Scientific Papers – Series Management, Economic Engineering in Agriculture and Rural Development, 23*(3), 863–870.

Sathyanarayanan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research, 27*(4S), 4023-4031. https://doi.org/10.53555/AJBR.v27i4S.4345

Sayıncı, B., Kara, M., Ercişli, S., Duyar, Ö., & Ertürk, Y. (2015). Elliptic Fourier analysis for shape distinction of Turkish hazelnut cultivars. *Erwerbs-Obstbau, 57*(1), 1–11. https://doi.org/10.1007/s10341-014-0221-7

Shafi, U., Mumtaz, R., Qureshi, M. D. M., Mahmood, Z., Tanveer, S. K., Haq, I. U., & Zaidi, S. M. H. (2023). Embedded AI for wheat yellow rust infection type classification. *IEEE Access, 11*, 23726–23738. https://doi.org/10.1109/ACCESS.2023.3254430

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35*(5), 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture, 161*, 272–279. https://doi.org/10.1016/j.compag.2018.03.032

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks for Visual Recognition, 11*, 1–8.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*(1), 9. https://doi.org/10.1186/s40537-016-0043-6

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8697–8710).