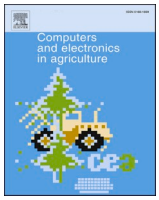
Computers and Electronics in Agriculture 237 (2025) 110491

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



A novel residual learning of multi-scale feature extraction model for the classification of rice grain varieties☆

Xudong Li [a], Yutong Wang [a], Happy Nkanta Monday [a,b,*], Grace Ugochi Nneji [a,b,*]

[a] School of International Education, Chengdu University of Technology Oxford Brookes College, China 610059
[b] Intelligent Computing Lab, HACE SOFTTECH, Lagos 102241, Nigeria

ARTICLE INFO

Keywords:
Ensemble model
Rice grain classification
Attention network
Explainability
Precision agriculture
Agricultural informatics

ABSTRACT

Rice serves as a fundamental food source for 50% of the world's population highlighting its crucial role in ensuring food security. Deep learning has become crucial tools in automating the labor-intensive task of rice grain classification, using digital image processing to evaluate quality and grain variety. This work utilized a large dataset consisting of 75,000 images of five different rice grain varieties. There are 15,000 images for each type, which capture distinct texture, form, and color features. Image augmentation approaches, such as normalization and transformations, are utilized to enhance model robustness and mitigate overfitting. The study presented a novel ensemble model that combined a customized attention mechanism with modified residual learning and multi-scale feature learning of parallel filters networks to improve the ability of features extraction and classification of rice grain varieties. A wide range of performance criteria is employed to assess the effectiveness of the model. The ensemble model demonstrated outstanding competence in classification tasks, achieving accuracy values close to 99%. The Grad-CAM visualization validates the model's attention towards pertinent characteristics among different rice grain varieties. The ensemble model outperformed pre-trained models and other works in terms of loss, accuracy, and F1-score, as shown by comparative analysis. This study enhances the field of agricultural informatics by boosting the accuracy of rice grain classification and food quality in general.

1. Introduction

Rice is a globally significant crop due to its crucial role in ensuring food security. Rice is a crucial staple for more than half of the world's population, especially in areas like Asia and Africa (Jabeen et al., 2023). Various environmental elements, such as temperature, humidity, and soil conditions, which fluctuate across geographical areas, can influence the quality and features of rice (Jin et al., 2022). Thus, accurately identifying different rice grain varieties is essential for achieving maximum crop productivity and quality.

Traditionally, the classification of rice grain varieties (Rajalakshmi et al., 2024) depended on visual inspection and evaluation of morphological characteristics and grain properties. The features encompass color, shape, taste, and aroma. Farmers and professionals commonly utilize their knowledge and discernment to distinguish between different rice grain varieties based on these characteristics (Zhang et al., 2018). Nevertheless, this approach might be influenced by personal opinions and can take a significant amount of time, which may result in errors during the identification procedure. Although conventional techniques for distinguishing rice grain varieties are important, they have certain limitations. The impact of individual viewpoints and the labor-intensive characteristics of these approaches emphasize the necessity for more accurate and efficient methodologies to ensure consistency and productivity of rice cultivation on a global scale.

In recent years, researchers have increasingly focused on sustainable agricultural practices to address challenges in crop production and classification. For instance, the adoption of black biodegradable mulching not only enhances crop yields but also reduces carbon emissions and improves environmental sustainability (Lin et al., 2024). Such practices highlight the importance of integrating advanced methods to optimize crop quality and productivity while mitigating environmental impacts. Similarly, genetic advancements, such as the identification of the OsTTG1 gene regulating anthocyanin biosynthesis in rice, emphasize the role of molecular mechanisms in improving rice traits (Yang et al.,

☆ This article is part of a special issue entitled: 'Agricultural Cybernetics' published in Computers and Electronics in Agriculture.

* Corresponding authors at: OBU Computing, Chengdu University of Technology, China 610059.
E-mail addresses: happy.monday@zy.cdut.edu.cn (H.N. Monday), grace.nneji@zy.cdut.edu.cn (G.U. Nneji).

https://doi.org/10.1016/j.compag.2025.110491
Received 29 January 2024; Received in revised form 23 February 2025; Accepted 29 April 2025

2021). These insights underscore the need for precise and scalable classification models that bridge traditional agricultural practices with modern technological innovations. Another researcher (Pan et al. 2024), utilized multi-modal feature extraction such as chemical composition, genetic markers, and environmental parameters into the classification process and this method enabled the development of more comprehensive models capable of processing multi-modal data for a holistic assessment of rice quality.

Many literature reviews have been conducted for binary classification; amongst them include Sun et al. 2014 who suggested the use of Support Vector Machine (SVM) method on a dataset with two classes, comprising 1700 data points related to rice grains. The model yielded a classification accuracy of 98.5 %. An independent analysis is performed on 200 data points from sixteen classes, yielding an accuracy of 87.16 % using the Support Vector Machine (SVM) technique (Liu et al., 2016). The manual classification of rice grains can be a laborious and costly procedure, as it mainly depends on human intervention. The evaluation process in manual methods may differ due to the limitations imposed by the evaluators' competence. In the field of machine learning and deep learning, characteristics are commonly classified using algorithms such as artificial neural network (ANN), support vector machine (SVM), logistic regression (LR), deep neural network (DNN), and Convolutional Neural Network (CNN). These algorithms have been studied and referenced by various researchers (Ebrahimi et al. 2014; Shrestha et al., 2016; Sabanci et al., 2017; Kaya & Saritas, 2019; Cortes & Vapnik, 1995; LaValley 2008; Dahl et al. 2013; Liu et al., 2017; Lin et al. 2018; Ahmed et al., 2020). In recent years, many digital image components have been used to evaluate the categorization and excellence of rice. These characteristics encompass several measurements, such as geometric dimensions like length and perimeter, fracture rate, whiteness, and the identification of cracks in rice grains. Image processing systems can extract numerous features of grain products. In addition, the manual process of making rapid conclusions can present difficulties when carrying out assessments on a large scale (Patrício & Rieder 2018). Rice, obtained from cereal crops, is an extensively produced and universally consumed food product. Rice is assessed and assigned value in the market depending on many features. Aukkapinyo et al. (2019) discuss parameters such as texture, form, color, and fracture rate. After acquiring the digital images of the products, several machine learning algorithms are utilized to determine these characteristics and perform categorization tasks. Machine vision technologies that employ image processing offer advantages over traditional methods that rely on manual labor (Barbedo 2016).

Additionally, some researchers emphasized the advantage of attention mechanism which makes the model concentrate on the distinct features of the image. Jing et al. (2023) suggested an improved DenseNet network and applied a channel attention mechanism squeeze-and-excitation to boost the intrinsic features. This model achieves an average accuracy classification of 99.4 %. Stephen et al. (2023) discuss the classification of rice leaf using four different CNN architectures, aside from the feature extractions using the CNN model, self-attention mechanism improves the selection process. Lastly, the ResNet34 with Self-attention achieved the best performance with 98.54 % accuracy. Tang et al. (2023) designed a 3D ResNet using hyperspectral image with 3D attention module for the extraction of distinct spectra features. This model achieved an overall accuracy of 97.47 %. Wang et al. (2021) presented an attention-based Depthwise separable neural network with Bayesian optimization based on MobileNet architecture. With different hyperparameter tweaking, cross-validation and the four classes of the dataset, the model achieved its test accuracy as 94.65 %. Chen et al. (2021) proposed a lightweight attention network with a pertained MobileNet-V2 and the loss function is optimized. The average classification accuracy achieved is 98.48 % for identifying rice plant diseases.

With all the researches carried out, there are some limitations with respect to robust feature extraction as the ensemble model utilizes the collective capabilities of single models, resulting in enhanced overall performance and increased accuracy, thereby alleviating the weaknesses of single models. On the other hand, a single stand-alone model may not efficiently capture all the intricacies and patterns in the data. Also for this specific dataset, the application of attention mechanism cannot be under-emphasized as this reduces computational cost by allowing the model to focus on intricate feature of each image. By incorporating the attention module into the fused models, these novel image classification techniques have the potential to produce favorable outcomes in identifying different rice grain varieties. This enables the model to allocate modified and optimized weights to each feature map channel, thereby decreasing the number of trainable parameters and creating a lightweight model. Hence, we are driven to create a novel framework that combines spatial attention with a residual-based learning ensemble model, with the aim of surpassing current methods and enabling more efficient categorization of rice grain varieties. The purpose of this study is to examine different rice grain varieties and develop a precise method for automatically identifying them using ensemble learning. This paper presents a novel ensemble approach for the classification of rice grain varieties. The successful completion of this study has the potential to enable farmers to make well-informed decisions on cultivation, leading to a more streamlined process of rice grain classification.

The significant contributions of this study are as follows:

- The study incorporates a customized attention mechanism to emphasize important characteristics in images of rice grains. The Spatial Attention Block, which surpassed the Squeeze-and-Excitation (SE) attention in first tests, showcases the model's capacity to concentrate on important textural and shape characteristics that are vital for achieving accurate categorization.
- To the best of our knowledge, the study is the first to introduce a novel ensemble model that effectively integrates the residual learning of multi-scale feature extraction of parallel filter networks in a synergistic manner. This innovative method enables more effective extraction of hierarchical features and analysis of features at several scales, resulting in a substantial improvement in the classification performance of the model. Additionally, the extraction of these features is more robust with the dual proposed model when compared to single models.
- Grad-CAM visualization is employed to interpret the focal regions of the model in rice grain classification, marking the first instance of its utilization in this context. These visualizations offer clear and transparent insights into the decision-making process of the model, guaranteeing that the learned patterns correspond to significant agronomic features.
- A comprehensive range of performance indicators, such as loss, accuracy, ROC-AUC, recall, precision, sensitivity, specificity, and F1 scores, are utilized to evaluate the model. This comprehensive evaluation framework guarantees that the model's effectiveness is rigorously assessed across multiple parameters of precision and dependability.
- The research undertakes a comprehensive comparative analysis using existing models and published literature. The ensemble model proposed exhibits exceptional performance in all parameters, hence establishing a new benchmark in rice grain classification.

The structure of the remaining parts of this study is as follows: Section II describes the materials and methodology. The proposed model architecture is discussed in the Second III while Section IV discusses the experimental results. Section V outlines the conclusion of this study, while Section VI provides a detailed limitation and future direction of this study.

2. Materials and methods

2.1. Dataset

The study utilized image collection of five distinct rice grain varieties, namely Arborio, Basmati, Ipsala, Jasmine, and Karacadag obtained from Kaggle Repository which is a public database. This collection is the first image dataset published by (Koklu et al., 2021), which has 75,000 grain images, with an equal distribution of 15,000 images for each species. This dataset displays distinct attributes of rice grains, such as its texture, shape, and color, which are essential for discerning between different rice grain varieties. Fig. 1 shows the different varieties of rice grains.

The images are collected under controlled laboratory conditions (Koklu et al., 2022). An Ikegami CCD imaging sensor (2.2 MP, resolution: 2048 × 1088 pixels) is used for capturing the images, which are taken within a black enclosed box measuring 14 cm × 18 cm. The camera is placed 15 cm above the samples, and uniform illumination is provided by an LED light source mounted at the top of the box. The consistent imaging conditions and the black background helped isolate the rice grains and highlight their distinct visual features.

Arborio is compact, robust, and elliptical species of rice grain, known for its firm texture while basmati is characterized by its elongated and slender shape, with a subtle curvature. Ipsala exhibits an oblong to slightly elongated shape, characterized by a smooth and consistent texture, Jasmine rice grains is characterized by its shorter, plumper and subtle pearlescent appearance and lastly, Karacadag has a robust, gently rounded form but may be slightly elongated and pointed.

In order to enhance the robustness and generalization capability of the proposed ensemble model, a wide range of data preprocessing and augmentation techniques are utilized. These techniques are essential for enhancing the model's performance by generating a more diverse training dataset from the existing images. The value of each pixel in the image is scaled by a factor of 1.0/255. The normalization step ensures that the pixel values fall within the range of [0, 1], which aids in achieving faster convergence during training by standardizing the input data and a rotation of 45 degrees is applied. This augmentation introduces random rotations to the images within 45 degrees, which aids in making the model insensitive to the orientation of objects in the images. A zoom range of 0.2 is employed, enabling the images to be randomly magnified or reduced by a maximum of 20 %. This transformation facilitates the model's ability to learn and identify items of different sizes. A shear transformation with a magnitude of 0.2 is applied. Shearing is the process of tilting the shape of the image, and this modification enhances the model's ability to withstand aberrations in the input images. The values for both the width and height shifts are adjusted to 0.2. This implies that the images have the potential to be randomly displaced both horizontally and vertically by a maximum of 20 % of the image's dimensions. This shift augmentation facilitates the model's ability to learn objects even in instances where they are not precisely positioned at the center. Both vertical and horizontal flips are enabled. This augmentation introduces random mirroring of the images along the vertical and horizontal axes, respectively. It is especially beneficial for enhancing the variety of the training set by offering distinct viewpoints of the objects. The original image dimensions of 250x250 pixels are resized to 50x50 pixels. The process of resizing minimizes the computing burden and memory needs for the model, allowing for quicker training and inference while preserving crucial recognition properties. Through the implementation of these preprocessing and augmentation approaches, the dataset is significantly enlarged, resulting in an enhanced capacity of the model to generalize to unseen data. This extensive augmentation technique ensures that the model is trained on a diverse range of image alterations, resulting in a more resilient and dependable performance in real-world scenarios. Additionally, for the evaluation of the proposed model, the study takes 70 % for training set and 30 % is evenly divided between the validation and test sets as presented in Table 1.

Despite the extensive preprocessing and augmentation techniques employed to enhance data diversity and model robustness, certain limitations related to the dataset's diversity and potential biases remain. Specifically, the dataset is derived from controlled conditions, including consistent lighting, uniform backgrounds, and aligned rice grain orientations. While these conditions ensure high-quality image acquisition, they may not fully reflect real-world scenarios where lighting, backgrounds, and grain orientations vary significantly. Such biases might impact the model's generalization to highly variable real-world datasets. Nevertheless, the augmentation techniques applied in this study significantly contribute to mitigating these limitations, enabling the model to generalize effectively across a wide range of image variations.

2.2. Proposed model

This subsection describes the basic model architecture, which forms the basis of our proposed method. The core building blocks include convolutional layers, batch normalization and ReLU activations. Subsequently, we present the optimization modules which include the residual learning block, multi-scale kernel learning structure and spatial attention module. These components are combined into an ensemble model to enhance classification performance.

2.2.1. Residual block structure

This architecture consists of a tailored Residual network presented in

Table 1
Dataset split across the five categories.

Rice Grain Variety	Train set (70 %)	Validation set (15 %)	Test set (15 %)
Arborio	10,950	2,250	2,250
Basmati	10,950	2,250	2,250
Ipsala	10,950	2,250	2,250
Jasmine	10,950	2,250	2,250
Karacadag	10,950	2,250	2,250
Total:	**52,500**	**11,250**	**11,250**

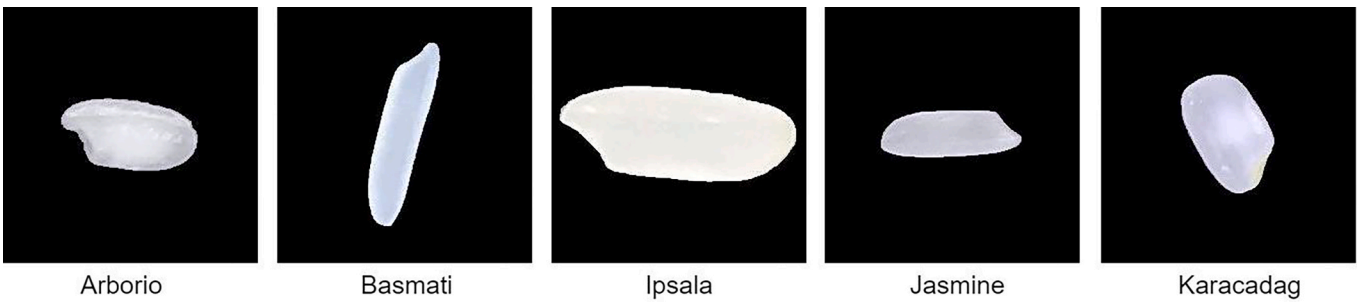


Fig. 1. Visual representation of the different rice grain varieties in the dataset.

upper view in Fig. 2, which incorporates several residual blocks as shown in Table 2. Every block comprises convolutional layers, batch normalization, and ReLU activation. After each residual block, a dropout layer is incorporated to mitigate overfitting and manage different degrees of feature abstraction. The residual connections incorporate conditional convolution and batch normalization to maintain alignment of feature dimensions, which is essential for enhancing the training stability and performance of the model. The central component of the model is built using a loop that sequentially goes through a set of stages, each consisting of a specific quantity of residual blocks. This design is adaptable, enabling the modification of the quantity of remaining blocks at each stage. This versatility allows for the modification of the architecture in terms of depth and complexity, making it suited for various task-specific requirements.

Fig. 2 is a variant of convolutional neural network with the utilization of skip to bypass certain layers. The previous input refers to the data that is received from the preceding block. The convolution layers comprise many convolution layers that utilize various kernel sizes such as 1x1 and 3x3 to extract features from the input. Batch normalization is applied in the features to normalize the output of the convolution layers in order to enhance the stability and efficiency of the neural network. Rectified Linear Unit (ReLU) activation function is used to incorporate non-linearity into the neural network, enabling it to acquire and understand intricate patterns.

The output of the last convolution layer is combined with the original input using the add method. The residual block is crucial for training extremely deep networks. A regularization method known as dropout that mitigates overfitting by stochastically excluding units from the neural network during the training process is introduced. A spatial attention network is introduced to allow the network focus on important features across spatial dimensions.

The residual block is stacked upon each other to form a complete stem of the proposed residual model as depicted in Fig. 2 while Table 2 shows the architectural parameters of the full stem residual model with skip connections. The fundamental equation that represents the layers in the residual block is shown in Eq. (1).

$$O = f(W^{*}X + b) \tag{1}$$

Where O denotes the output, f is the activation function (ReLU), W represents the weight matrix of the kernel, X represents the input matrix,

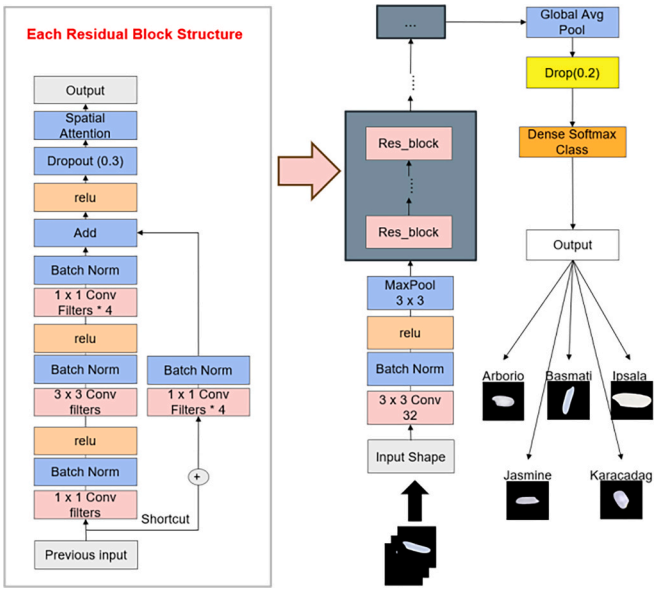


Fig. 2. Architecture of the proposed residual learning model of skip connections.

b represents the bias term, and * symbolizes the convolution process. The output after applying batch normalization is calculated from Eq. (2).

$$O_{BN} = \gamma\left(\frac{O - \mu}{\sigma}\right) + \beta \tag{2}$$

where μ and σ are the mean and standard deviation of the input while γ and β are the learnable parameters of the normalization. The additional operation that establishes the residual connection is given in Eq. (3).

$$O_{res} = O_{conv} + X \tag{3}$$

The output of the convolution operation, denoted as O_{conv}, is added to the input X to obtain the residual output O_{res}. The inclusion of these elements and processes in residual networks is to effectively acquire the ability to learn identity functions, which helps to circumvent the issue of the vanishing gradient problem that often arises in deep networks.

2.2.2. Attention mechanism

In order to improve the model's ability to recognize features, two attention mechanisms are taken into account; Squeeze-and-Excitation (SE) and Spatial Attention Block. During initial testing, it is noticed that the Spatial Attention Block performed better than SE, potentially because of the dataset's properties, which showed more prominent variations in shape and texture rather than variances in color. Therefore, the Spatial Attention Block is solely included in the experimental framework, as depicted in Fig. 3. This block performs average and maximum pooling on the input feature maps, combines these pooled features, and subsequently utilizes a convolutional layer to produce an attention map.

The map is standardized using a sigmoid function and then multiplied with the original feature map element by element to emphasize important spatial regions, hence strengthening the model's attention on crucial picture information and improving classification accuracy. Fig. 3 illustrates the Spatial Attention Block with the purpose of directing the neural network's attention towards specific spatial characteristics in the input data whereas Table 3 depicts the overall architectural parameters of the spatial attention block.

The input shape refers to the initial dimensionality of the data being fed into the attention block. The model includes two types of pooling layers, specifically max pooling and average pooling. Each of these operations decreases the number of dimensions in the feature maps, highlighting distinct characteristics. The max pooling layer prioritizes the most prominent features, while average pooling takes into account the overall arrangement. The results of these pooling layers are subsequently merged along the channel axis. This process merges the unique viewpoints of the feature maps into a unified picture. The convolution layer with a 7×7 kernel size applies a convolution operation which is used to combine the features across the channels, enabling the network to selectively emphasize certain aspects. The output is a feature map that the network utilizes to assess the significance of various spatial points in the input data. The overall structure can be represented mathematically as presented in Eq. (4).

$$O_{ij} = \sum_{k} (I_{ij} * W_k) + b \tag{4}$$

Where O_{ij} is the output feature map obtained by convolving the input feature map I_{ij} with the weight matrix W_k for the convolutional filter, index k is used to refer to the input channels, and a bias term b. The summation is performed across all input channels, and the operation is executed at each spatial position (i, j). The max and average pooling methods are straightforward; they extract the maximum and average values from the input feature map within a defined window, respectively. The utilization of this attention mechanism enables the network to concentrate on the most informative segments of the input.

Table 2
Architectural parameters of the full stem residual learning model of skip connections.

Layer Name	Layer Type	Filter Size	Stride	Padding	Output Channel	Activation Function
conv1	Conv2D + BN	3 x 3	2	same	32	relu
max_pool	MaxPooling2D	3 x 3	2	same	32	–
Residual block Structure (6x)						
conv1	Conv2D + BN	1 x 1	strides	same	filters	relu
conv2	Conv2D + BN	3 x 3	1	same	filters	relu
conv3	Conv2D + BN	1 x 1	1	same	filters*4	relu
shortcut	Conv2D + BN	1 x 1	strides	same	filters*4	–
add (special case)	Add	–	–	–	–	relu
dropout	Dropout	–	–	–	–	–
sa_block	sa_block	–	–	–	–	sigmoid
Global Average Pool						
global_pool	GlobalAverage Pooling2D	–	–	–	–	–
dropout	Dropout	–	–	–	–	–
dense	Dense	–	–	–	5	softmax

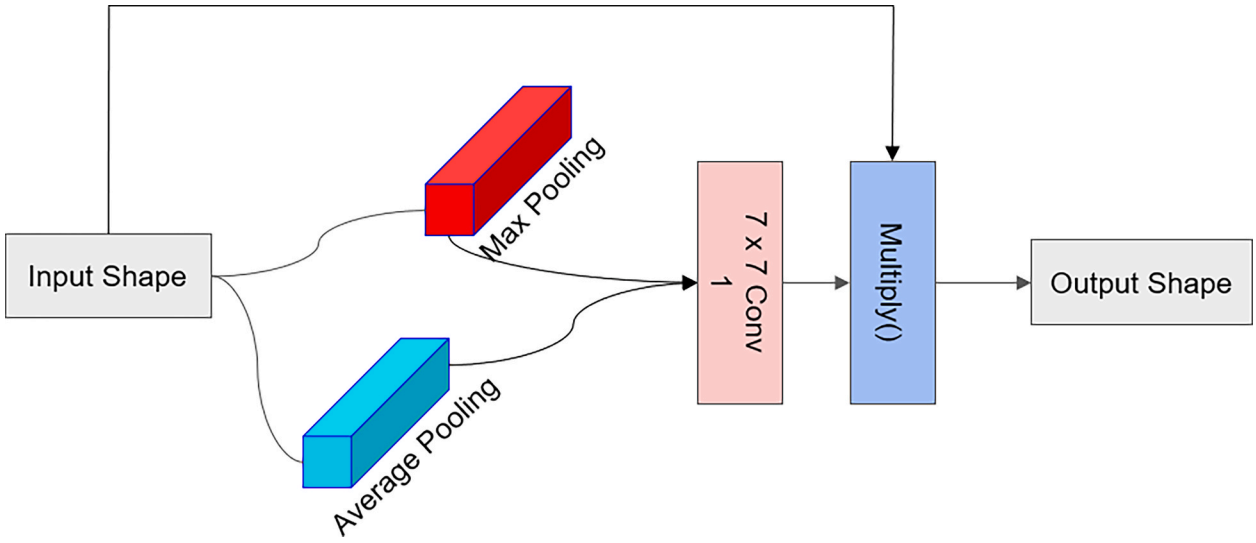


Fig. 3. Spatial Attention Block.

Table 3
Architectural parameters of the spatial attention block.

Layer Name	Layer Type	Filter Size	Stride	Padding	Output Channel	Activation Fuction
avg_pool	Lambda (mean)	–	–	–	1	–
max_pool	Lambda (max)	–	–	–	1	–
concat	Concatenate	–	–	–	2	–
conv	Conv2D	7 x 7	1	same	1	sigmoid
multiply	Multiply	–	–	–	–	–

2.2.3. *Multi-Scale kernel learning architecture*

The model utilizes a multi-scale kernel learning network, as depicted in the upper view in Fig. 4. This network consists of several multi-scale kernel blocks, as detailed in Table 4. The blocks are organized using convolutional kernels of different sizes, allowing for the simultaneous extraction of information at different scales. The design of the architecture enables the simultaneous processing of intricate local patterns and extensive contextual information from the images. In order to enhance the model's ability to generalize and prevent overfitting, a dropout layer is incorporated after each multi-scale kernel block. This design decision guarantees that the network takes advantage of a strong feature extraction method that is resistant to the subtleties of the input data.

Fig. 5 depicts a customized constituent of the multi-scale structure engineered to handle data at various scales concurrently. The structure has various convolutions with varied kernel sizes 1*x*1, 3*x*3, and 5*x*5. The 1*x*1 convolutions also function to decrease dimensionality, a method referred to as bottleneck, prior to the bigger 3*x*3 and 5*x*5 convolutions in order to manage computational resources effectively. The introduction of the spatial attention layer is to enhance the multi-scale kernel module to prioritize the important spatial characteristics in the input data, potentially enhancing its capacity to identify intricate patterns in the images. The results of all convolution operations are combined together along the channel dimension, enabling the network to gain knowledge from a wide range of features. The equation that governs the convolution processes in the multi-scale kernel blocks is given by Eq. (5).

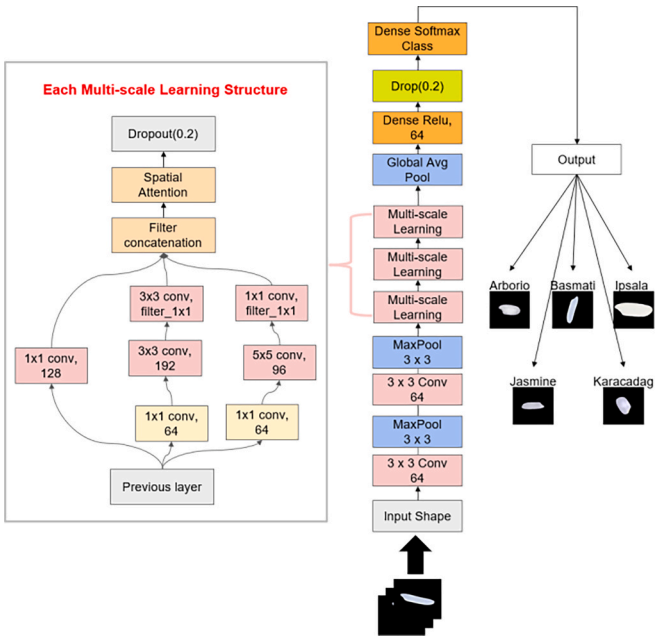


Fig. 4. Proposed architecture of the multi-scale learning structure.

$$F_{ij}^{l} = \sigma\left(\sum_{m=1}^{M_{l-1}}\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1} K_{pq}^{lm}.F_{i+j,j+q}^{l-1} + B^{l}\right) \tag{5}$$

The equation represents the calculation of the activation F_{ij}^{l} in the multi-scale block. It involves the summation of the product of filter weights K_{pq}^{lm} and the activation values $F_{i+j,j+q}^{l-1}$ from the previous layer, together with a bias term, B^{l}. The result is then passed through the sigmoid function σ to obtain the final activation value. Where F represents the feature map at layer l, σ denotes the ReLU activation function, K represents the convolution kernel, B represents the bias term, M represents the number of input feature maps, and P, Q represent the dimensions of the kernel. The operation of the filter concatenation combines the feature maps generated by various filters 1x1, 3x3, and 5x5 convolutions along the depth dimension by concatenation. The given expression is presented in Eq. (6).

$$F_{concat} = Concat(F_{1x1}, F_{3x3}, F_{5x5}) \tag{6}$$

The concatenated feature map, denoted as F_{concat}, is obtained by concatenating the feature maps F_{1x1}, F_{3x3}, and F_{5x5}. Spatial attention is applied to concentrate on particular spatial regions of the input. The process usually entails calculating attention scores and subsequently applying these scores to the input feature map as denoted in Eq. (7).

$$A_{ij} = AttentionFunction(F_{ij}), where F_{ij}^{'} = A_{ij}.F_{ij} \tag{7}$$

The value of A_{ij} is determined by applying the *AttentionFunction* to (F_{ij}). The derivative of $F_{ij}^{'}$ with respect to i, j is equal to the product of A_{ij} and F_{ij}. Dropout is applied to randomly set some of the input characteristics to zero in order to mitigate the risk of overfitting. The mathematical formulation of dropout is given in Eq. (8).

$$F_{ij}^{'} = F_{ij}.d_{ij} \tag{8}$$

where d_{ij} is the random variable of 0 with probability p (dropout rate) and the value 1 with probability $1-p$. The full stem of the multi-scale architecture is constructed by stacking these blocks upon one another as presented in Fig. 4 while the parameters of the full stem multi-scale model are presented in Table 4

2.2.4. The ensemble model

The ultimate model design deviates from conventional ensemble techniques like voting or weighted averaging, which are usually better suited for integrating separate model outputs at the decision level. Alternatively, this method utilizes a decision-level fusion strategy by combining the outputs of the residual learning and multi-scale kernels models using concatenation method. This approach leverages the hierarchical feature abstraction skills of skip connections and the multi-scale features extraction. The concatenated output vector is subjected to additional processing, which includes a fully connected layer and a dropout layer, in order to efficiently combine and enhance the feature information from both networks. The ultimate classification is executed by employing a softmax layer.

It should be noted that the residual learning which utilizes skip connections enables gradients to bypass one or more layers, hence enhancing the efficiency of back-propagation. This approach effectively

Table 4
Architectural parameters of the full stem multi-scale model.

Layer Name	Layer Type	Filter Size	Stride	Padding	Output Channel	Activation Function
conv1	Conv2D	3 x 3	2	same	64	relu
max_pool1	MaxPooling2D	3 x 3	2	same	64	relu
conv2	Conv2D	3 x 3	1	same	32	relu
max_pool2	MaxPooling2D	3 x 3	2	same	32	relu
Multi-scale block (3x)						
branch1_conv1	Conv2D	1 x 1	1	same	128	relu
branch2_conv1	Conv2D	1 x 1	1	same	64	relu
branch2_conv2	Conv2D	3 x 3	1	same	192	relu
branch2_conv3	Conv2D	1 x 1	1	same	filter_1x1	relu
branch3_conv1	Conv2D	1 x 1	1	same	64	relu
branch3_conv2	Conv2D	5 x 5	1	same	96	relu
branch3_conv3	Conv2D	1 x 1	1	same	filter_1x1	relu
concatenate	Concatenate	–	–	–	–	–
sa_block	sa_block	–	–	–	–	sigmoid
dropout	Dropout	–	–	–	–	–
Global Average Pooling						
global_pool	GlobalAverage Pooling2D	–	–	–	–	–
dense	Dense	–	–	–	64	relu
dropout	Dropout	–	–	–	–	–
dense	Dense	–	–	–	5	softmax

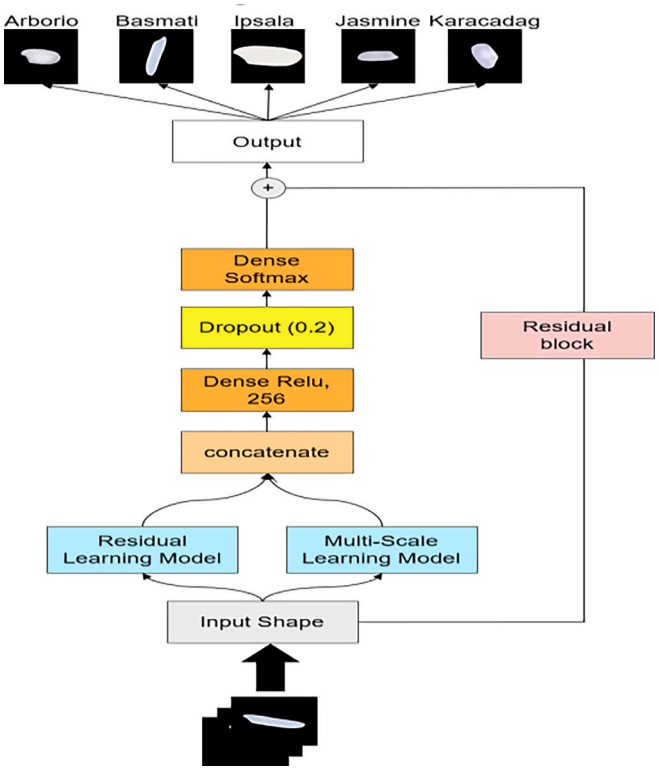


Fig. 5. Architecture of the proposed ensemble model.

addresses the issue of the vanishing gradient problem, allowing for the training of deeper networks without any decline in performance. Skip connections in residual networks expedite the convergence of the model during training by facilitating a direct pathway for the flow of gradients. As a consequence, this leads to enhanced training speed and potentially greater precision. More so, residual connections, through the use of identity mapping, assist in retaining information from previous layers. This facilitates the learning of residual functions and enhances the overall representational capacity of the model. The multi-scale kernel architecture enables the network to concurrently handle information at different scales. This feature allows the model to accurately capture intricate details as well as wider contextual characteristics, hence enhancing its performance on intricate task.

By employing various filter sizes simultaneously, the model can acquire more resilient and all-encompassing feature representations. This is especially advantageous since the images have diverse spatial hierarchies. By using bottleneck layers that utilize 1x1 convolutional filters, the total number of training parameters is greatly reduced. The reduction is accomplished by initially compressing the feature maps using 1x1 convolutions, followed by the application of bigger convolutional filters. This process effectively reduces the computational load. The 1x1 convolutional filters act as effective layers for projecting features, allowing the input feature space to be transformed into a space with less dimensions. This modification decreases the computational complexity while preserving the fundamental attributes of the data. The model achieves dimensionality reduction by utilizing 1x1 convolutions, resulting in improved control over overfitting and enhanced generalization. This technique ensures that the model stays computationally viable, even when faced with the increased intricacy of multi-scale learning. The proposed model combines the benefits of residual learning and multi-scale kernel architecture to construct a neural network that is both robust and efficient, while also achieving excellent performance. The inclusion of skip connections in residual learning enhances the flow of gradients and improves training efficiency. Additionally, the utilization of multi-scale learning further boosts the extraction of features. Utilizing bottleneck techniques alongside 1x1 convolutional filters enhances the model's optimization by decreasing the number of training parameters and computing cost. This ensures that the advantages of increased complexity are attained without imposing excessive computational requirements. Table 5 details the architectural parameters of the ensemble model.Table 6 shows the parameters of the excection environment.

Fig. 5 depicts the proposed ensemble model, which encompasses the input layer as the initial stage where the images enter the network. The image undergoes first changes, such as normalization, scaling, or augmentation. The individual model performs feature extraction to extract features from the images. The outputs from the branches are merged using a fusion technique of concatenation. The final prediction of the ensemble model which is expected to classify the rice grain images into one of the following varieties; Arborio, Basmati, Ipsala, Jasmine, and Karacadag. The proposed ensemble technique exploits the capabilities of each individual model, minimizing their particular limitations. The fundamental concept of an ensemble model can be expressed mathematically as given in Eq. (9).

$$P(y|x) = \frac{1}{N}\sum_{i=1}^{N} P_i(y|x) \tag{9}$$

The expression in Eq. (9) represents the average of the conditional probabilities $P_i(y|x)$ over a set of N values, where $P(y|x)$ is the ensemble's predicted probability of class y given input x. N is the number of models in the ensemble.

In addition to the proposed ensemble model, we compared its performance against several state-of-the-art models including ResNet50, VGG16, EfficientNetB0 and DenseNet201. These pre-trained algorithms are widely used for image classification problems, and their comparison will reveal the improvement presented by the multi-scale and residual learning techniques of our proposed model in enhancing image classification. The analysis and interpretation of the comparative performance of these models will be presented in the results section.

2.3. Evaluation indicators

To thoroughly assess the effectiveness of the created model in categorizing rice grain varieties, a range of performance evaluation measures are utilized in this work. These metrics offer both a quantitative assessment of the model's prediction accuracy and insights into its strengths and limitations across different areas, such as its capacity to handle intricate data. The subsequent subsections provide a comprehensive explanation of the definition, significance, and computation techniques for each assessment metric. The performance metrics have been calculated using the data of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confusion matrix is utilized to assess the classification effectiveness of the suggested algorithm.

2.3.1. Loss

This metric quantifies the discrepancy between the anticipated values of the model and the actual values. A tailored Cross-Entropy Loss function is employed, with Y_{true} being the actual label and Y_{pred} denoting the anticipated label. This adaptation is tailored to more effectively meet the specific demands of this study as presented in Eq. (10).

$$L(Y_{true}, Y_{pred}) = \frac{1}{N}\sum_{i=1}^{N}\left(-\sum Y_{true} \bullet \log(clip(Y_{true}, \epsilon, 1-\epsilon))\right) \tag{10}$$

$L(Y_{true}, Y_{pred})$ denotes the loss function, which measures the dissimilarity between the true labels Y_{true} and the Y_{pred}. $\frac{1}{N}$ calculates the average by dividing the loss by the number of samples N in the dataset. $\sum_{i=1}^{N}$ represents the summation of the loss over all N instances in the collection. In a one-hot encoded vector, the value 1 represents the correct class and

Table 5
Architectural parameters of the proposed ensemble model.

Layer Name	Layer Type	Filter Size	Stride	Padding	Output Channel	Activation Function
resnet_output	resnet_model	–	–	–	5	–
multi-scale_output	multi-scale_model	–	–	–	5	–
concatenated	Concatenate				10	
dense	Dense	–	–	–	256	relu
dropout	Dropout	–	–	–	–	–
final_dense	Dense	–	–	–	5	softmax

Table 6
Parameters of the execution environment.

Hardware Unit	Specifications
Central Processing Unit	Intel i7 10750H 2.59 GHz
RAM	16 GB
Graphic Card	Nvidia RTX 2070
Operating System	Windows 10
Programming Language	Python 3.10
Framework	Tensorflow 2.9
Batch Size	32
Learning Rate	1e-5 –> 1e-3 (5 epoch stage up) –> 1e-3 (keep 5 epoch) –> 1e-4 –> 1e-5…
Iteration	40 (max) (Accuracy stops after 3 epochs when it stops rising)

0 represents otherwise. log is used to convert the predicted probability into a loss value. $clip(Y_{true}, \epsilon, 1-\epsilon)$ is the clip function employed to avoid calculating the logarithm of zero, which is undefined. It guarantees that the predicted probability Y_{pred} is kept at a minimum value greater than 0 by a small ε, hence preventing the occurrence of infinite loss values. The negative symbol, $''-''$ preceding the summation signifies summation of the negative logarithm of probabilities. This is a common convention used to define cross-entropy loss. The loss becomes positive since the logarithm of values between 0 and 1 is negative.

2.3.2. Accuracy

Accuracy is the ratio of accurately predicted samples to the total number of samples in a model. It offers a precise assessment of the model's overall prediction efficacy. This study utilizes accuracy metric that takes into account the argument of the highest probability in both the actual and predicted labels. This ensures a comparison of the most likely categorical results, as described in Eq. (11).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

2.3.3. ROC-AUC curve

This curve evaluates the model's performance at various thresholds. The model's effectiveness in differentiating between classes increases as the area under the curve becomes larger.

2.3.4. Confusion matrix

This matrix helps to identify associations between the classifier accuracy and the test outcomes of the test, providing detailed summary of the correct and incorrect categorizations for the rice grain varieties with the quantitative values for true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

2.3.5. Recall

Recall is a quantitative measure that assesses the model's capacity to accurately identify samples belonging to the positive class. It quantifies the proportion of actual positive samples that are correctly identified by the model, as shown in Eq. (12).

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{12}$$

2.3.6. Precision

Precision measures the accuracy of the model in predicting positive class samples correctly. The measure is crucial for evaluating the precision of the model's positive class predictions, as defined in Eq. (13).

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{13}$$

2.3.7. Specificity

Specificity quantifies the model's ability to accurately detect samples belonging to the negative class. It offers a perspective on the model's ability to identify occurrences that are irrelevant, as described in Eq. (14).

$$Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositive} \tag{14}$$

2.3.8. Sensitivity

Sensitivity, also known as recall in some contexts, measures the model's ability to correctly identify positive class samples. It is a critical metric in evaluating the model's effectiveness in detecting positive cases, as defined in Eq. (15).

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{15}$$

2.3.9. F1 score

The F1 Score is a mathematical measure that combines accuracy and memory to assess the overall effectiveness of a model. It is particularly useful when achieving a balance between precision and recall is crucial, as described in Eq. (16).

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{16}$$

2.3.10. Precision-Recall curve

This curve illustrates the relationship between precision and recall at different levels. It is particularly advantageous for assessing the effectiveness of models on datasets with intricate features.

2.3.11. Grad-CAM

Grad-CAM is an interpretive methodology that identifies and emphasizes significant regions within images that play a critical role in a model's decision-making process. It aids in comprehending the key aspects and decision-making processes of the model in visual activities.

2.4. Hardware Specifications

2.4.1. Experimental results

In this study, the model pre-training employs a custom loss function and an accuracy calculation method. The custom loss function aims to ensure that the predicted probabilities remain within a numerically stable range and optimizes the model by calculating cross-entropy with respect to the true labels. The accuracy function measures the model's

accuracy by comparing the predicted labels with the true labels. To comprehensively evaluate the model's performance, this study utilizes an array of metrics, including loss, accuracy, ROC-AUC curve, recall, precision, PR curve, sensitivity, specificity, confusion matrix, F1-score, and precision-recall curve. Particularly, Grad-CAM analysis is conducted on the two models prior to their ensemble in this study. The purpose of this analysis is to identify key areas in the models when processing different types of rice grain images, thereby determining which model requires further adjustment and optimization to enhance the overall performance of the final ensemble model. 4.1 Model Performance of the Ensemble Model.

Fig. 6(a) illustrates the loss graph, which demonstrates efficient learning. The training and validation loss exhibit a significant decline during the initial epochs, suggesting a rapid acquisition of knowledge. The training loss exhibits a consistent and gradual decline, whilst the validation loss reaches a plateau, indicating that the model is starting to generalize without suffering from overfitting. The near convergence of the two lines at the conclusion of the training indicates that the model is finely adjusted and sustains its performance on unfamiliar data, which is a favorable indication of its predictive capacity. The accuracy curve showcases the model's ability to consistently achieve high accuracy on both the training and validation sets, hence confirming its capacity for generalization, as illustrated in Fig. 6(b).

The training and validation accuracy exhibit a pronounced initial surge, suggesting rapid acquisition of knowledge and enhancement in the model's capacity to accurately identify the training data. As the epochs advance, the accuracies of both models reach a plateau, indicating that the model has attained a high level of proficiency approximately 99 %. The close alignment of the training and validation accuracy towards the end indicates that the model has a strong ability to generalize and is likely to regularly perform well on new, unseen data. The model's high accuracy level is a strong indication of its effectiveness. The ROC and PR curves present a perfect AUC of 1.00 for all classes, signifying the model's excellent discrimination ability with no ambiguity between the classes. Fig. 7(a) represents the ROC curve, which evaluates the performance of the proposed model. The ROC curve is a visual depiction of a classifier's performance, illustrating the relationship between the true positive rate (TPR) and the false positive rate (FPR) across different threshold values. The AUC quantifies the model's capacity to differentiate between different rice grain classes. Each line corresponds to each rice variety being categorized, and all of them have AUC value of 1.00. An AUC value of 1 indicates flawless classification, where there is a 100 % TPR and a 0 % FPR for all thresholds. The model demonstrates exceptional discriminative capability for each class and can effectively differentiate between the various rice grain kinds without any mistakes.

Fig. 7(b) depicts the PR curve which is used to visually show the relationship between precision and recall at various thresholds. From the curve, it is evident that the lines for all the classes are tightly clustered in the upper right corner. This suggests that the model consistently achieves high levels of precision and recall across all classes. In essence, this model demonstrates the ability to reliably distinguish the positive class while minimizing the occurrence of both false positives and false negatives, which is a characteristic of an exceptional classifier. The proximity of these curves to the upper right corner indicates the model's reliability in its predictions.

The confusion matrix in Fig. 8 (comes from the test set) shows the model's performance across several rice grain varieties. The cells along the diagonal (from top left to bottom right) represent the count of accurate predictions for each class, which are remarkably high, suggesting a robust true positive rate. For example, Karacadag has made 2,243 accurate predictions with minimal misclassifications, primarily as Arborio. Basmati, Ipsala, and Jasmine exhibit a high number of true positives, with 2,246, 2,225, and 2,250 accurate classifications, respectively, with no misclassifications for Jasmine. Arborio shows slight ambiguity in distinguishing between different classes, but it consistently achieves a high true positive rate, accurately predicting 2,225 instances, indicating the model's overall strong performance.

2.4.2. *Statistical analysis*

Table 7 presents the performance metrics of the proposed ensemble model for classifying five different rice grain varieties. The model demonstrates exceptional precision and recall for the Karacadag variety, achieving 99.82 % and 99.69 %, respectively. These values indicate that the model is highly accurate in identifying Karacadag, with an incredibly low incidence of false positives and false negatives. The F1-score of 99.76 % highlights a well-balanced trade-off between precision and recall. For the Basmati variety, the model shows slightly higher precision (99.91 %) and recall (99.82 %) compared to Karacadag, suggesting an even lower occurrence of misclassifications. The F1-score of 99.87 %

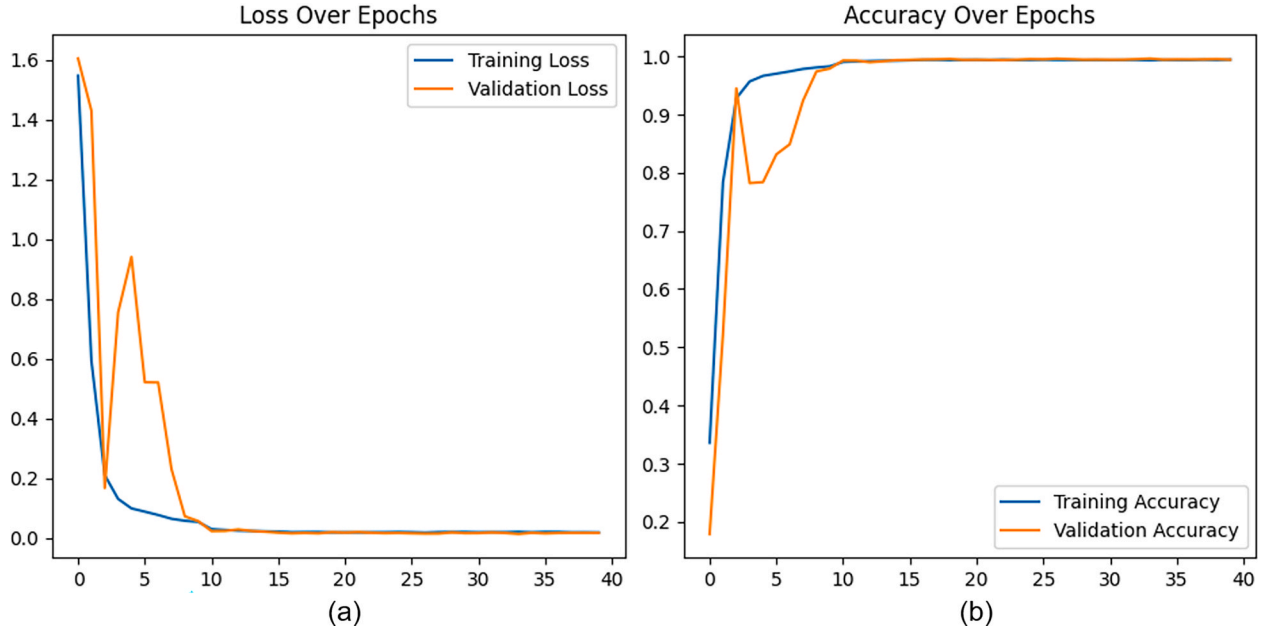


Fig. 6. Training and validation performance on loss and accuracy metrics for the ensemble model.

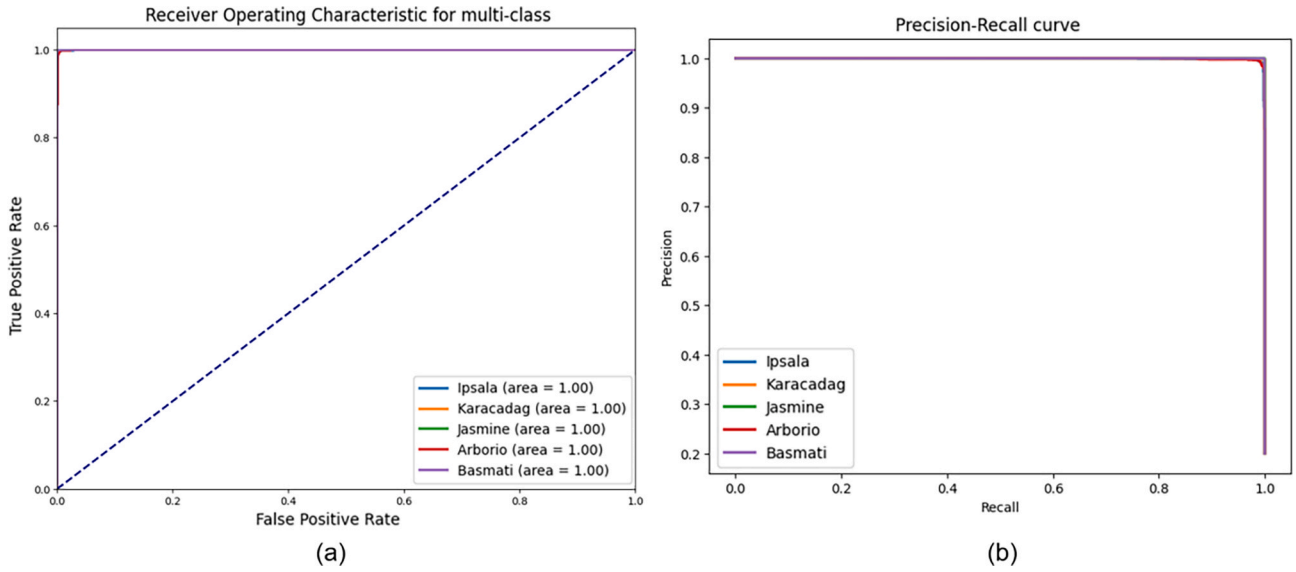


Fig. 7. (a) ROC-AUC and (b) PR Curve for the ensemble model.

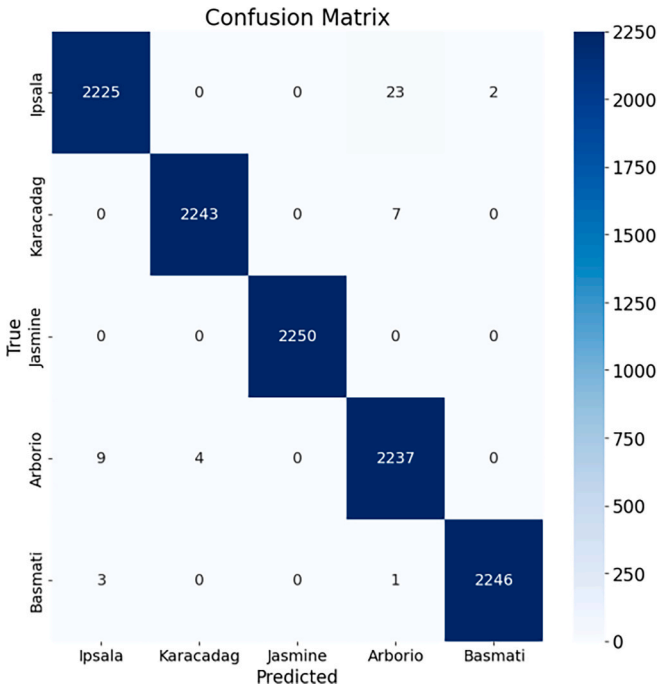


Fig. 8. Confusion Matrix of the ensemble model (The data presented here is derived from the test set).

Table 7
Performance of the ensemble model across the different rice grain varieties.

	Ipsala	Karacadag	Jasmine	Arborio	Basmati
TP	2,225	2,243	2,250	2,225	2,246
FN	25	7	0	13	4
FP	12	4	0	31	2
TN	8,988	8,996	9,000	8,969	8,998
Precision	99.46 %	99.82 %	100 %	98.63 %	99.91 %
Recall	98.89 %	99.69 %	100 %	99.42 %	99.82 %
F1-score	99.18 %	99.76 %	100 %	99.03 %	99.87 %
Sensitivity	98.89 %	99.69 %	100 %	99.42 %	99.82 %
Specificity	99.87 %	99.96 %	100 %	99.66 %	99.98 %

confirms this model's robust performance in identifying Basmati. In the case of Ipsala, the model achieves 99.46 % precision and 98.89 % recall, with an F1-score of 99.18 %. While precision is slightly lower than Karacadag and Basmati, the model still performs admirably. The performance for Arborio variety shows a precision of 98.63 % and recall of 99.42 %, with an F1-score of 99.03 %. While the precision is marginally lower, recall remains high, indicating the model's ability to correctly identify most instances of Arborio. Finally, the Jasmine variety achieves perfect results with 100 % precision, recall, and F1-score, reflecting the model's flawless classification for this rice type. The model's specificity is similarly high across all varieties, further validating its strong performance in distinguishing between different rice grains.

Sensitivity and Specificity are performance metrics used to evaluate the model's ability to detect positive and negative examples. Sensitivity corresponds to recall, while Specificity reflects the model's ability to accurately identify negative examples. The scores for each class exhibit exceptional levels, approaching 1.0 (100 %), signifying that the model is remarkably proficient in both detecting positive occurrences and accurately identifying negative examples. This indicates that the model possesses a high level of discriminatory power for each type of rice grain and demonstrates resilience in accurately classifying them.

2.4.3. Explainability and visualization

This subsection showcases Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations of various rice grain varieties. Grad-CAM is a method used to visually represent the specific regions of the image that the proposed ensemble model with attention network is prioritizing while making predictions, as presented in Fig. 9 for classifying rice varieties. It produces visual explanations for the model's predictions, emphasizing the crucial areas in the input images that impact the output. The interpretability of the model provides understanding to the specific qualities that the model prioritizes, hence enabling better-informed judgments while refining the model. The visual maps generated by Grad-CAM have the ability to expose any biases that the proposed ensemble model may have acquired throughout the training process. To address biased emphasis areas, the model is retrained and the training procedure is modified to reduce these biases. Another benefit is the increase in the training efficiency of the proposed ensemble model by gaining insight into the specific areas that the model is prioritizing its attention on, which optimizes the training process.

The Grad-CAM analysis reveals that the model's attention is primarily directed towards the core section of Arborio rice grain when

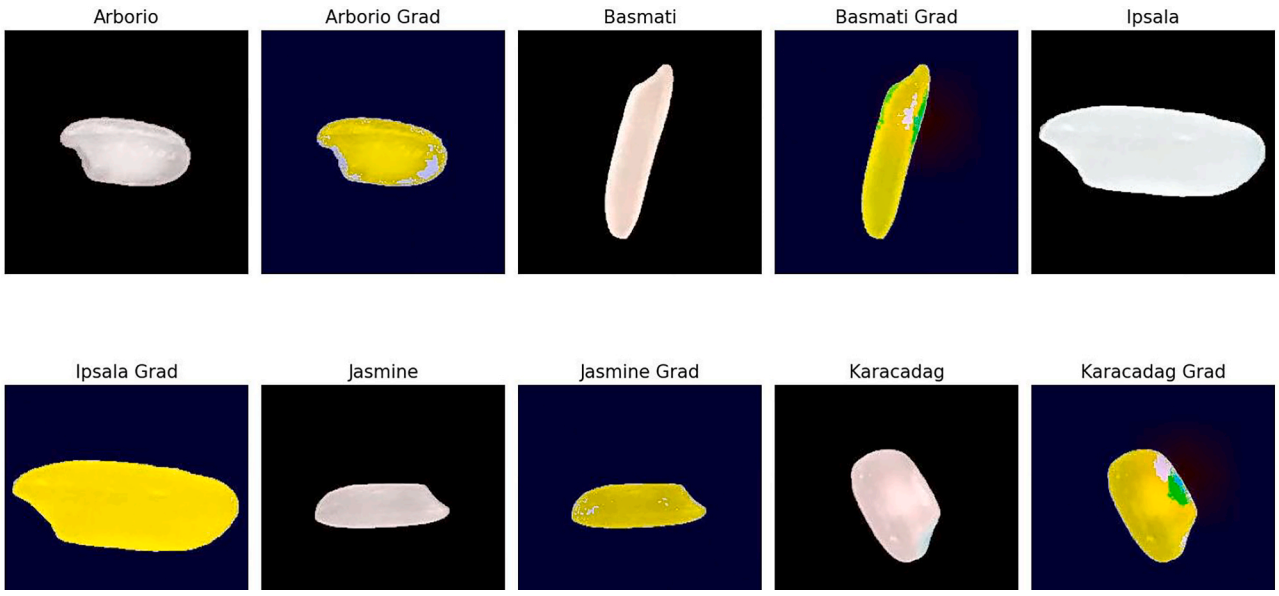


Fig. 9. Grad-CAM visualization for explainability of the proposed ensemble model.

making predictions. The image of Basmati rice grain shows that the model prominently emphasizes the grain's longitudinal axis, which is a distinctive feature of its elongated shape. The model places significant emphasis on the complete structure of the Ipsala rice grain, indicating that the model utilizes the grain's entire form for classification purposes. The highlighted region on Jasmine is primarily located in the center to lower portion of the grain, suggesting that these characteristics are unique to this category. The activation of Karacadag rice grain is focused on the middle section of the grain, similar to Arborio rice grain. This suggests that this characteristic plays a substantial role in the model's decision-making process. The Grad-CAM results not only validate the model's focus on biologically significant features but also offer interpretability that is valuable for stakeholders, such as agricultural experts. By providing clear visual explanations of the model's decisions, Grad-CAM facilitates confidence in its predictions, enabling stakeholders to better understand and trust the model's behavior. These visualizations also support practical applications by aligning the model's attention with the distinguishing morphological features of each rice grain variety. The Grad-CAM results offer a qualitative visualization of the areas where the model focuses its attention to make decisions. The model's attention is concentrated in the bright yellow regions, indicating the highest level of confidence in its prediction. This alignment with biologically significant features suggests that the model is learning the most relevant characteristics for precise predictions. The utilization of Grad-CAM visualization helps to analyze and understand the behavior of the model, leading to a favorable impact on both model development and practical use. Since the highlighted sections correlate to significant segments of the rice grains for the classification purposes, it implies that the model is acquiring suitable characteristics for accurate predictions. Additionally, the heatmap analysis provides further insights into the model's attention mechanism. For instance, areas with higher attention scores (represented in bright yellow) correlate with the features most indicative of each rice grain variety, enhancing the transparency of the model's decision-making process.

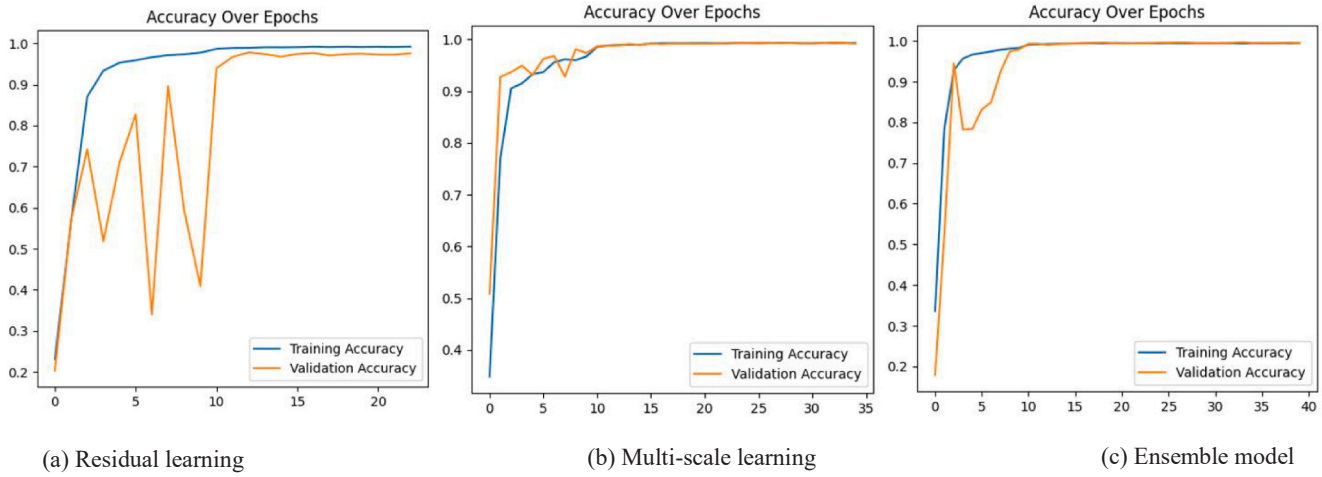
The Grad-CAM visualizations, combined with the statistical analysis and performance metrics, validate the ensemble model's ability to accurately classify rice grain varieties while maintaining a high level of interpretability. These findings not only demonstrate the model's robustness but also highlight its capacity to generalize effectively to unseen data. This alignment of the model's attention with biologically and morphologically significant features ensures that the predictions are not only precise but also meaningful for practical applications in agricultural quality control and variety classification.

2.4.4. Comparison of the ensemble model with the single models

Fig. 10 illustrates the accuracy comparison of the three models across training epochs. The residual learning model exhibits a rapid surge in training accuracy during the early epochs, quickly approaching its maximum accuracy presented in Fig. 10(a). The validation accuracy closely follows the training accuracy after an initial period of fluctuation, indicating that the model effectively generalizes without overfitting. Both training and validation accuracies eventually stabilize, signaling a consistent and steady learning process. Fig. 10(b) shows that the multi-scale learning model exhibits a rapid rise in training accuracy, similar to the residual learning model. However, the validation accuracy starts higher than the training accuracy from the outset, which could indicate an anomaly, such as an easier validation set or some other model-specific issue. Over time, the validation accuracy gradually aligns with the training accuracy, with both accuracies converging and stabilizing. While this convergence reflects strong generalization, the initial discrepancy and the slight delay in the validation accuracy catching up suggest potential challenges in the model's optimization or data handling. The ensemble model exhibits a rapid rise in training accuracy, quickly approaching near-perfect accuracy. The validation accuracy, following a minor deviation, converges with the training accuracy, indicating exceptional generalization. The ensemble model exhibits a continuously high level of accuracy over the remaining epochs presented in Fig. 10(c). All three models exhibit convergence of training and validation accuracies to high levels, generally surpassing 90 %. This suggests that the models are effectively learning and capable of making accurate predictions on new, unseen data. The ensemble model appears to be the most promising, since it exhibits the least discrepancy between training and validation accuracy, furthermore, the validation accuracy does not exceed the training accuracy in the early stages of training, avoiding anomalies in multi-scale learning models. Fig. 11 illustrates the loss comparison of the three models across training epochs.

The residual learning model in Fig. 11(a) exhibits an initial sharp decrease in training loss, indicating a high rate of learning. The validation loss exhibits a consistent decreasing trend with occasional fluctuations, ultimately reaching a stable state. The loss for both training and validation steadily decreases to nearly zero, indicating a successful reduction in error as the learning process advances.

The training loss for the multi-scale model as depicted in Fig. 11(b) exhibits an initial fast decline followed by a gradual convergence



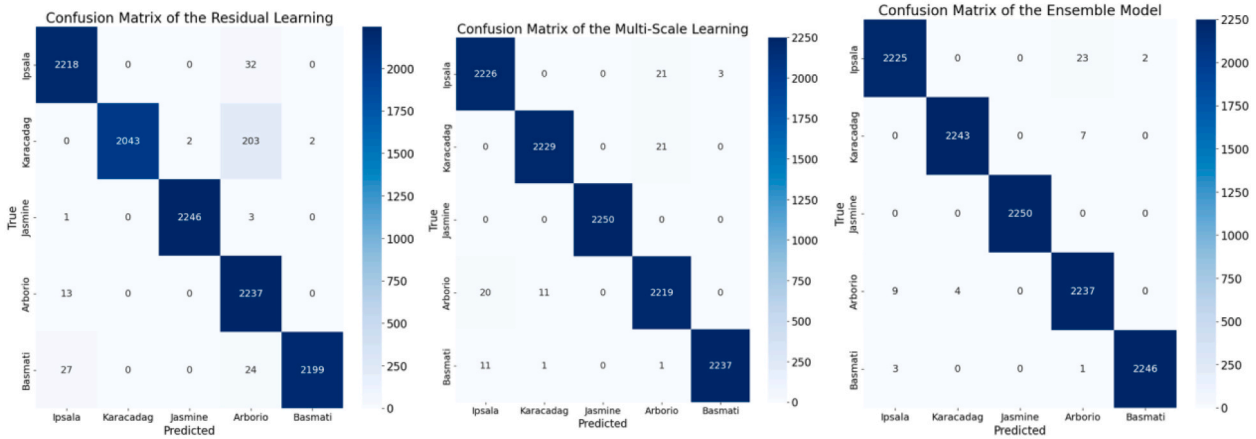
(a) Residual learning (b) Multi-scale learning (c) Ensemble model

Fig. 10. Accuracy comparison of the three models: (a) Residual learning, (b) Multi-scale learning, (c) Ensemble model.

(a) Residual learning (b) Multi-scale learning (c) Ensemble model

Fig. 11. Loss comparison of the three models: (a) Residual learning, (b) Multi-scale learning, (c) Ensemble model.

towards a low and stable value. The validation loss exhibits a similar pattern, initially decreasing and then stabilizing, indicating consistent learning without notable overfitting. Both lines reach a plateau, demonstrating a decrease in error rates as the epochs grow. This convergence, while indicating strong generalization, could benefit from further investigation due to the initial discrepancy. According to Fig. 11 (c), the ensemble model demonstrates a rapid decrease in training loss, reaching a low value nearly quickly. The validation loss exhibits a quick decline and roughly corresponds to the training loss, indicating that the model effectively generalizes. The ensemble model consistently



(a) Residual learning (b) Multi-scale learning (c) Ensemble model

Fig. 12. Confusion matrix comparison of the three models: (a) Residual learning, (b) Multi-scale learning, (c) Ensemble model.

maintains a low loss during the training period, furthermore, the validation loss does not lower than the training loss in the early stages of training, avoiding anomalies in multi-scale learning models, indicating a very effective learning process and minimal mistake.

The three models exhibit substantial reductions in training and validation losses, especially in the initial epochs, indicating effective learning of the models. The ensemble model is notable for its rapid convergence and minimal loss values, indicating that it is the most efficient model for minimizing error in both the training and validation datasets. Fig. 12 displays confusion matrices for three distinct learning models, utilized to demonstrate the classification performance on a test dataset with known true values. The residual learning model showed strong performance with many true positives as shown in Fig. 12(a), especially for the Jasmine class with 2,246 accurate predictions and very few misclassifications. However, there were some misclassifications, particularly in the Karacadag class, where 203 instances were incorrectly predicted as Arborio.Despite these misclassifications, the diagonal values representing accurate predictions remained high across all classes, suggesting that the overall performance of the model was very stable. The multi-scale learning model in Fig. 12(b) has excellent performance, achieving zero misclassifications for the Jasmine class and high true positive rates for the remaining classes. However, the model occasionally misclassifies Arborio as Ipsala and Basmati as Ipsala, resulting in a few off-diagonal values. In general, the diagonal values exhibit a high magnitude, indicating commendable performance. However, the error classification of Arborio and Ipsala is flawed compared to the residual learning models, while Karacadag and Jasmine have more concentrated errors. The confusion matrix of the ensemble model in Fig. 12(c) demonstrates flawless classification for the Jasmine class and minimal misclassifications for the remaining classes.

Among the three models, it exhibits the largest number of accurate positive predictions and the lowest number of incorrect classifications, particularly for Arborio and Karacadag, which are frequently more difficult to differentiate. The presence of a nearly flawless diagonal with low off-diagonal numbers indicates exceptional ability in accurately categorizing all classes. Collectively, the three models exhibit robust categorization capabilities. However, the ensemble model distinguishes itself as the most precise, showcasing the potential advantages of amalgamating diverse modeling techniques to enhance prediction efficacy. Fig. 13 displays the Grad-CAM visuals for three distinct models, delineating the specific regions that each model prioritizes while classifying various rice grain varieties. The visualizations from Fig. 13(a) demonstrate that the residual learning model focused attention on distinct regions of the rice grains. For example, the primary areas of interest on the Basmati and Ipsala grains are located along their longitudinal axes, which are significant distinguishing characteristics attributed to their shape. The depiction for Arborio, however, focuses on the overall structure, which is a distinctive feature of this type of grain. The visualization of the multi-scale learning model in Fig. 13(b) appears to emphasize wider regions of the rice grains in comparison to the residual learning model.

This implies that the model is considering a greater range of characteristics across the surface of the grains. The highlighted areas on Basmati and Ipsala exhibit a more dispersed distribution, suggesting an emphasis on both texture and form characteristics. The Ensemble model in Fig. 13(c) integrates the methodologies of the previous models, demonstrating concentrated yet inclusive attention across all aspects. The highlighted regions have a balanced distribution, indicating that this model is effectively taking into account both shape and textural characteristics. The visualizations suggest that this model has the ability to combine the advantages of residual and multi-scale learning, resulting in enhanced overall performance.

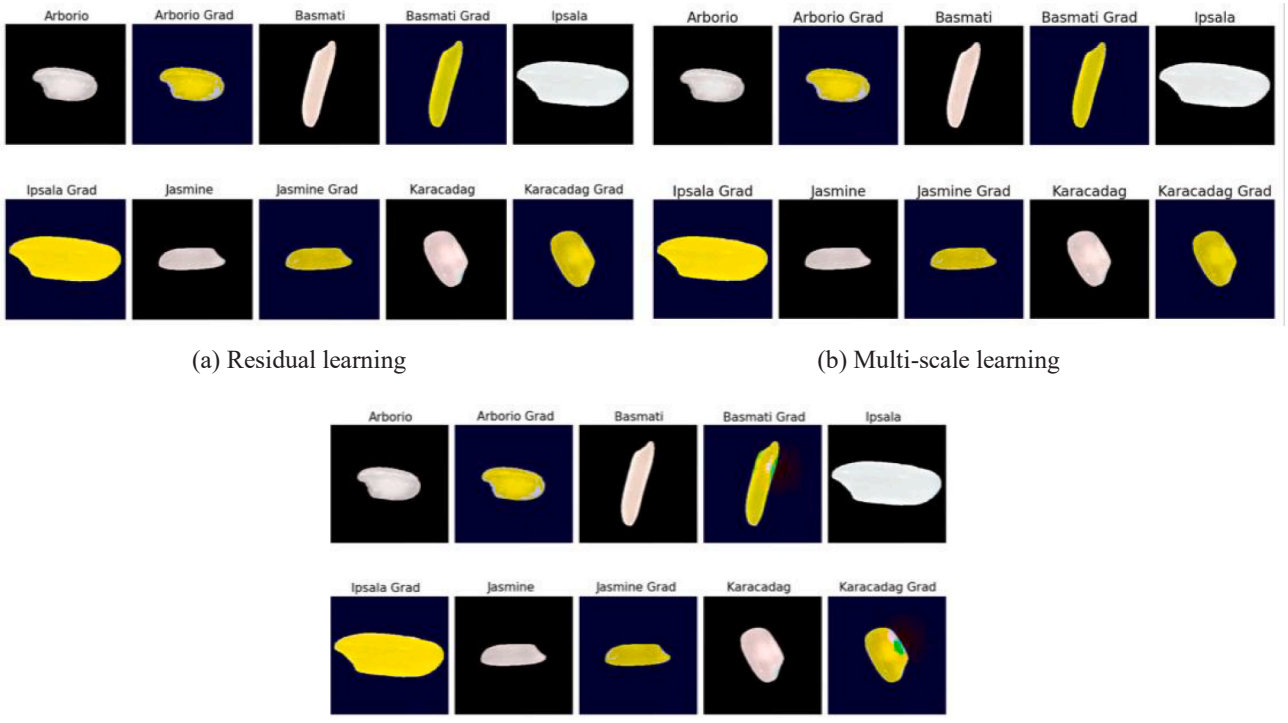


Fig. 13. Grad-CAM Visualization comparison of the three models: (a) Residual learning, (b) Multi-scale learning, (c) Ensemble model.

Grad-CAM visualizations provide insight into the specific areas of focus for each model while making predictions about the rice grains. The ensemble model demonstrates a high level of balance, which contribute to its higher performance, as indicated by the prior comparisons of confusion matrices. These visualizations are essential for comprehending the rationale behind specific predictions made by the models and are vital for improving and interpreting the models. The precision-recall comparison in Fig. 14 reveals that the residual learning, multi-scale learning, and ensemble model all exhibit similar and outstanding performance. Fig. 14(a) shows that the residual learning exhibits exceptional performance, as seen by the curves for each class label closely following the upper-right corner of the graph. This indicates that the model has a high precision, meaning it has a low rate of false positives, and a high recall, meaning it has a low rate of false negatives, for each class label. The multi-scale learning model presented in Fig. 14(b) like the residual learning model, exhibits precision-recall curves that are very identical to a perfect score. Additionally, the curves for all class labels are very close to the top-right corner. This signifies a significant level of model efficacy. The precision-recall curves of the Ensemble Model presented in Fig. 14(c) closely resemble those of the residual and multi-scale learning models. The curves for all class labels exhibit a close alignment with the top-right corner, indicating exceptional precision and recall.

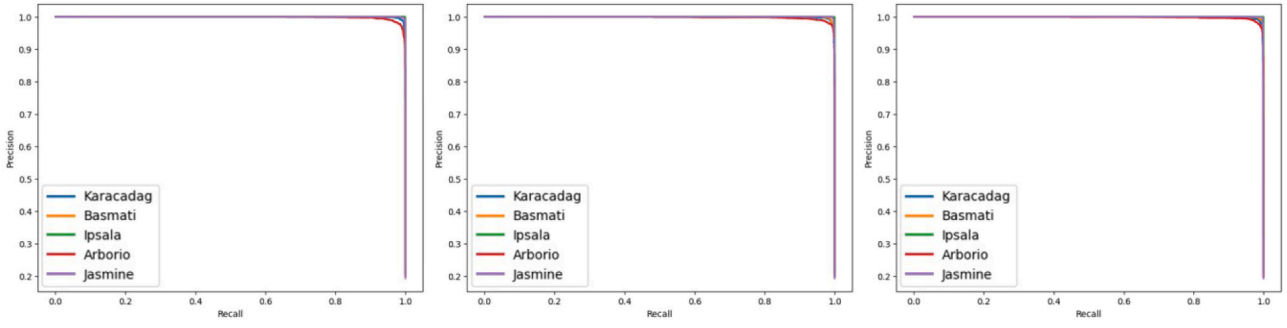
In summary, the precision-recall curves for all three models demonstrate their ability to accurately predict each class label with a high level of certainty and minimum mistakes. The curves indicate that there is minimal or no compromise between precision and recall for any of the models, demonstrating their strong performance in accurately classifying the various rice grain varieties. The comparison of ROC-AUC values among the three models, namely Residual Learning, Multi-scale Learning, and Ensemble Model, showcases the performance for each of the models as presented in Fig. 15. The residual learning model presented in Fig. 15(a) shows that the ROC curves for each class label closely follow the left-hand and top borders of the graph, suggesting a high AUC value near to 1. This implies that the residual learning model achieves a high true positive rate for each class while keeping the false positive rate low. The performance of the multi-scale learning model presented in Fig. 15(b) is also excellent, as evidenced by the curves for each class label closely aligning with the ideal line of an AUC of 1. This model showcases a remarkable equilibrium between sensitivity and specificity. The ensemble model maintains the same pattern, as evidenced by the ROC curves in Fig. 15(c) for each class label showing AUCs that are extremely close to the maximum value of 1. The curves for all class labels demonstrate its strong discriminative capacity to accurately classify the positive class at various thresholds. The ROC-AUC results indicate that all three models exhibit high efficacy in differentiating between the various classes of rice grain varieties, with minimal overlap between classes and exceptional overall classification performance.

Table 8 presents a performance comparison of the three distinct models: Residual Learning, Multi-scale Learning, and Ensemble Model. The Residual Learning model achieved a loss of 0.026, an accuracy of 99.14 %, a recall of 98.94 %, a precision of 98.96 %, a sensitivity (equivalent to recall) of 98.94 %, an F1-score of 98.95 %, and a specificity of 99.74 %. The Multi-scale Learning model, exhibits a slightly diminished loss of 0.024, accompanied with an accuracy of 99.16 %, which is marginally superior to the Residual Learning model. The recall rate is 99.08 %, the precision rate is 99.10 %, the sensitivity rate is 99.08 %, the F1-score is 99.09 %, and the specificity rate is 99.77 %. These values indicate a remarkably high level of performance. With a loss of 0.023, the Ensemble model demonstrates superior performance in accurately identifying the correct classes. The accuracy is at its peak with a value of 99.26 %, while recall and sensitivity also reach a high level of 99.15 %, and precision matches this value as well. The F1-score, recall, and precision are all similar at a rate of 99.15 %. However, specificity surpasses them all with a rate of 99.79 %. In summary, the Ensemble model has greater performance compared to the other two models in all criteria, suggesting that it effectively integrates the benefits of the Residual and Multi-scale Learning models. The Multi-scale Learning model exhibits superior performance compared to the Residual Learning model, particularly in terms of loss, accuracy, and specificity. Although the Residual Learning model lags somewhat behind the other two models, it still demonstrates exceptional performance metrics.

2.5. Ablation study

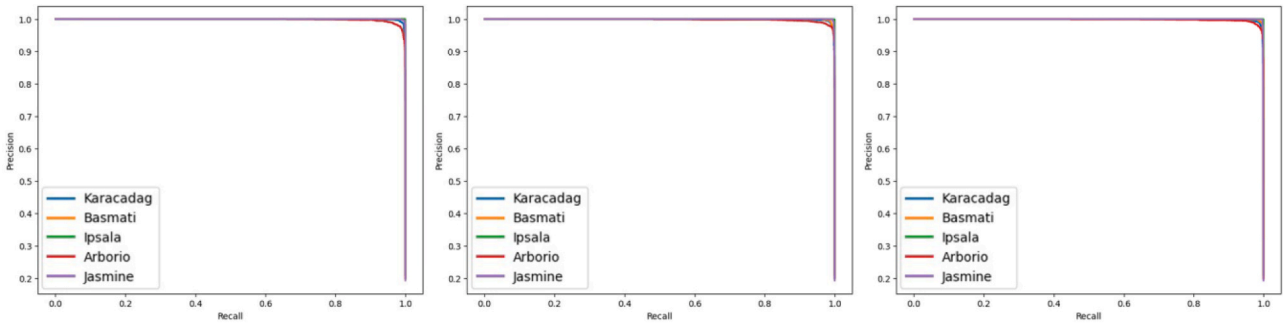
Before determining the final model configuration to achieve best performance, a sequence of hyperparameter experiments is carried out. These trials enhanced comprehension of the influence of each hyperparameter on the model's performance and enabled the model to be optimized to attain the currently reported levels of performance. During the initial trials, although there is evidence of learning, the performance measures are far lower compared to the findings currently being presented. The iterative modifications of these parameters resulted in the subsequent observations:

- Decreasing the learning rate caused the training process to become excessively slow in the first stages, while increasing the learning rate caused fluctuations to occur near the optimal solution. Equilibrium is achieved by balancing the speed of convergence and the stability of the model, therefore preventing overfitting through careful experimentation.
- Modifying the dropout rate had a substantial impact on the model's ability to generalize. It is noted that a modest rise in the dropout rate could alleviate overfitting. Nevertheless, an extremely high dropout rate could have a severe impact on the model's training effectiveness.
- Augmenting the batch size accelerated the training of the model, but it also required additional memory and had an effect on the accuracy



(a) Residual learning (b) Multi-scale learning (c) Ensemble model

Fig. 14. Precision-Recall comparison of the three models: (a) Residual learning, (b) Multi-scale learning, (c) Ensemble model.



(a) Residual learning (b) Multi-scale learning (c) Ensemble model

Fig. 15. ROC-AUC comparison of the three models: (a) Residual learning, (b) Multi-scale learning, (c) Ensemble model.

Table 8
Performance comparison of the three models: (a) Residual network, (b) Multi-scale network, (c) Ensemble model.

Model	Loss	Accuracy	Recall	Precision	Sensitivity	F1-score	Specificity
Residual learning	0.026	99.14 %	98.94 %	98.96 %	98.94 %	98.95 %	99.74 %
Multi-scale learning	0.024	99.16 %	99.08 %	99.10 %	99.08 %	99.09 %	99.77 %
Ensemble model	0.014	99.51 %	99.64 %	99.76 %	99.64 %	99.76 %	99.91 %

of the model. The optimal batch size is discovered by empirical investigation, enabling a balanced compromise between training efficiency and model performance.

The following part will provide a comprehensive analysis of how these hyperparameters impact the performance of the model, supported by specific experimental data.

2.5.1. Learning rate

The loss and accuracy plots of the ensemble model in Fig. 16 demonstrate a notable enhancement in the early stages of training, as the loss values begin at a high level and rapidly decrease to below 0.2, while the accuracy steadily rises from 20 % to over 80 %. Throughout the epochs, both metrics reach a stable state. However, the final loss on the validation set seems to level off at a larger value compared to the training set, suggesting potential overfitting. The ultimate accuracy values demonstrate a significant plateau for the training set at 95 %, with a marginally lower but consistent outcome for the validation set at 88 %, indicating the model's resilience while also emphasizing the possibility for enhanced generalization.

According to Fig. 17, the ensemble model demonstrates exceptional performance across all class labels, with both the ROC-AUC and Precision-Recall curves earning near-perfect scores. The ROC-AUC curve on the left shows that each class label, Karacadag, Basmati, Ipsala, Arborio, and Jasmine has an AUC near to 1.0. This indicates a high true positive rate and a low false positive rate across different thresholds, which is considered excellent and the high AUC values indicate the model's extraordinary capacity to differentiate between classes.

The precision-recall curve on the right demonstrates consistently good precision for all recall levels in each class, converging towards the upper right corner. This indicates a combination of high accuracy and strong recall. The curves indicate that the model has a high accuracy in

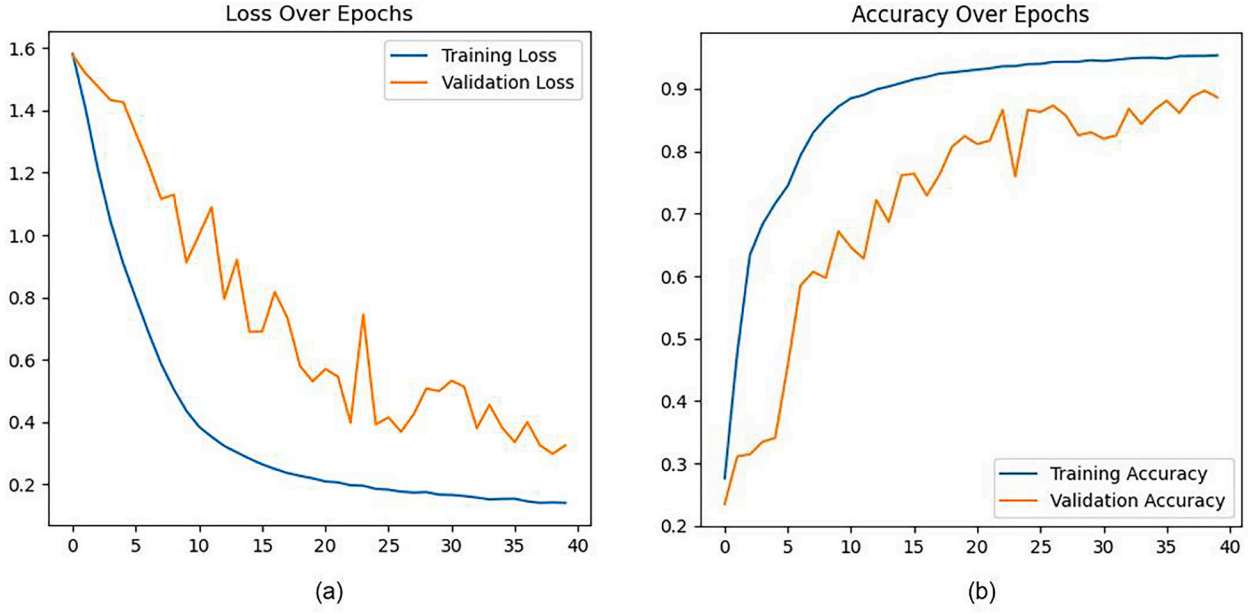


Fig. 16. Loss and accuracy plots of the ensemble model for the learning rate of 0.0001.

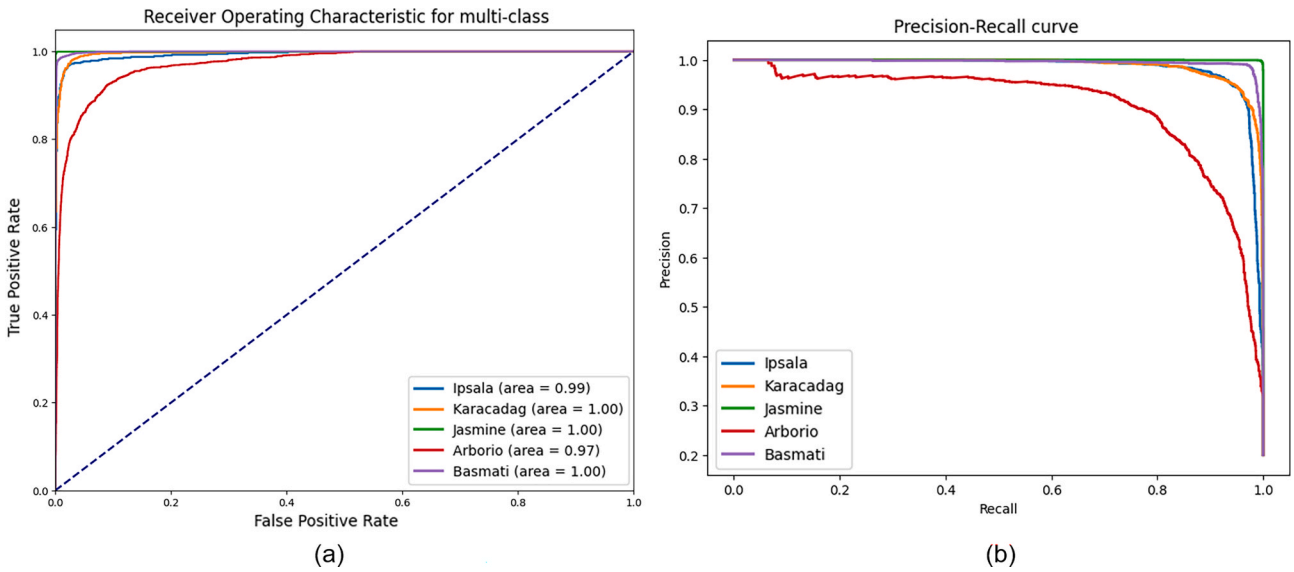


Fig. 17. ROC-AUC and precision-Recall curves of the ensemble model for the learning rate of 0.0001.

properly identifying positive instances of each rice grain variety, with a low occurrence of both false positives and false negatives.

The Grad-CAM visualizations presented in Fig. 18 demonstrates the ensemble model's performance at a learning rate of 0.0001 by showing the regions of interest that had the greatest impact on the classification judgments for each rice grain variety. The model's attention is mostly directed towards the central and upper-middle parts of the Arborio rice grains, which are recognized for their robust and oval form. The visualizations of Basmati rice grains demonstrate focus along the grains' length, which is in line with Basmati's distinctive long and slim shape. The model takes into account the complete surface area of the Ipsala grains, suggesting that both the form and texture throughout the grain play a key role in classification. The Grad-CAM analysis reveals that the model's attention is primarily directed towards the middle region of the

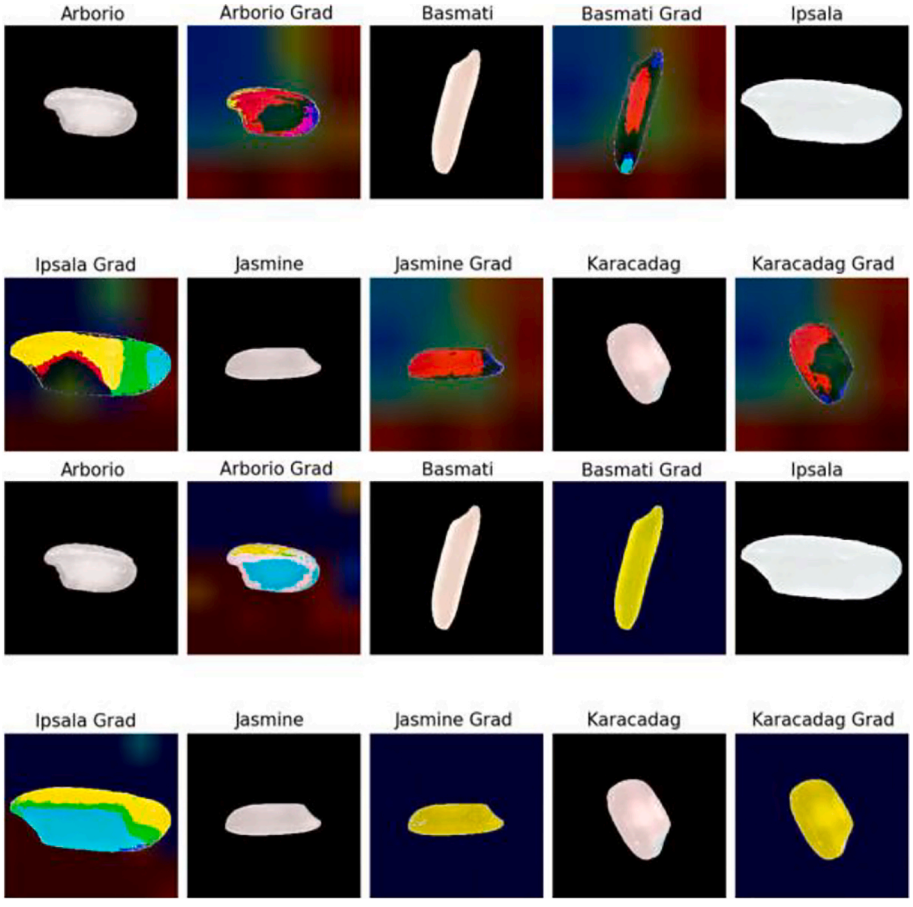


Fig. 18. Grad-CAM visualization of the ensemble model for learning rate of 0.0001.

grains, which aligns with Jasmine's characteristic shorter and plumper form in comparison to Basmati rice grains. The regions that stand out on the grains of Karacadag are mainly found in the central area, which could be related to the distinctive features of its shape. The confusion matrix depicted in Fig. 19 showcases the classification efficacy of the ensemble model with a learning rate of 0.0001 for each type of rice grain. The model exhibits robust performance by accurately categorizing 2,241 occurrences of Karacadag, with a minimal amount of misclassifications as other categories. Similarly, the model demonstrates high accuracy with Basmati rice, correctly identifying 2,244 instances, with only 4 misclassifications as Ipsala and 2 as Jasmine. For Jasmine, the model correctly identifies 2,214 cases, though 32 instances are incorrectly classified as Arborio, suggesting some overlap in characteristics between Jasmine and Arborio. The model encounters challenges with Arborio, accurately categorizing 1,511 cases, but incorrectly classifying 666 instances as Karacadag. This suggests that there may be shared characteristics between Arborio and Karacadag that lead to confusion in classification.

Also, the model accurately identifies 1,739 instances of Ipsala rice grains, but it also makes a significant number of misclassifications as Basmati, indicating that there may be difficulties in discriminating between these two types.

Table 9 presents the performance metrics of the rice grain classification model, using a learning rate of 0.0001, a dropout rate of 0.3, and a batch size of 32. Precision ranges from 77.09 % to 99.95 %, with Jasmine achieving the highest precision at 99.95 %, indicating no false positives. Karacadag shows the lowest precision at 77.09 %, implying a higher number of false positives. Recall rates vary, with Ipsala scoring 77.29 %, while Basmati stands out with a higher recall of 99.73 %, showing fewer false negatives. The F1-score, which balances precision and recall, is highest for Jasmine at 99.17 % and lowest for Arborio at 78.05 %.

2.5.2. Dropout rate

The loss and accuracy plots for the ensemble model, which incorporate a dropout rate of 0.2, provide a visual representation of the model's learning performance as presented in Fig. 20. The training loss drops sharply at the outset, indicating swift model adaptation. The validation loss initially follows this downward trend but later fluctuates with occasional spikes, pointing to difficulties in generalizing unseen data. Despite these variations, it eventually stabilizes and aligns closely with the training loss. The training accuracy rises rapidly at first, reaching nearly 100 % and plateauing throughout the rest of the epochs. The validation accuracy mirrors this upward trend, suggesting strong generalization capabilities. Overall, the ensemble model demonstrates a high level of proficiency, achieving stable and near-perfect accuracy, indicative of successful learning and generalization.

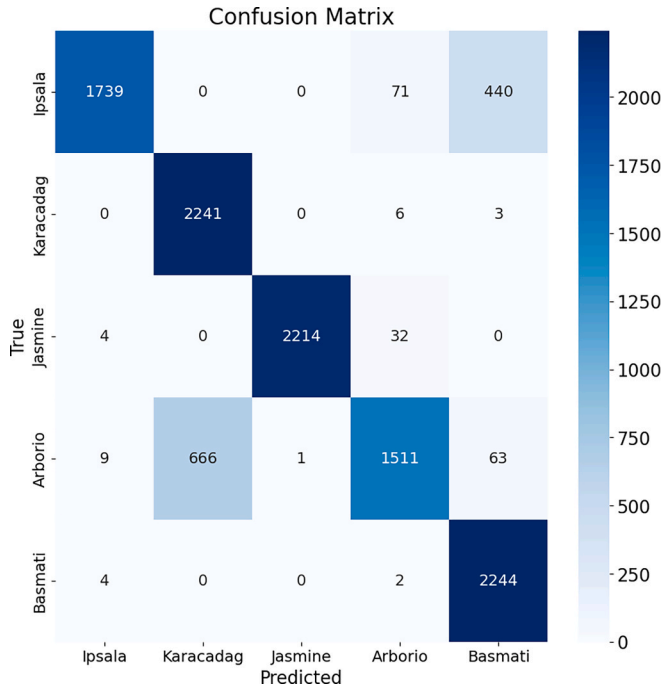


Fig. 19. Confusion matrix of the ensemble model for learning rate of 0.0001.

Table 9
Performance of the ensemble model across the different rice grains varieties using a learning rate of 0.0001.

	Ipsala	Karacadag	Jasmine	Arborio	Basmati
TP	1,739	2,241	2,214	1,511	2,244
FN	511	9	36	739	6
FP	17	666	1	111	506
TN	8,983	8,334	8,999	8,889	8,494
Precision	99.03 %	77.09 %	99.95 %	93.16 %	81.6 %
Recall	77.29 %	99.60 %	98.40 %	67.16 %	99.73 %
F1-score	86.82 %	86.91 %	99.17 %	78.05 %	89.76 %
Sensitivity	77.29 %	99.60 %	98.40 %	67.16 %	99.73 %
Specificity	99.81 %	92.60 %	99.99 %	98.77 %	94.38 %

The ensemble model, with a dropout rate of 0.2 for each class label, exhibits exceptional classification performance, as shown by the ROC-AUC and PR curves in Fig. 21. The Receiver Operating Characteristic curve shows that each class label, namely Karacadag, Basmati, Ipsala, Arborio, and Jasmine, has achieved a perfect AUC score of 1.00.

The model's flawless AUC score demonstrates its exceptional ability to accurately identify true positives while avoiding false positives, efficiently distinguishing between the various classes with a high level of certainty. The PR curve for all class labels approaches the upper right corner of the graph, indicating that the model consistently achieves high accuracy at various degrees of recall. The model repeatedly demonstrates the ability to accurately identify instances of the positive class while minimizing both false positives and false negatives. The Grad-CAM visualizations in Fig. 22 delineate the specific regions within each rice grain image that the ensemble model prioritized when ascertaining the class label. The Arborio model demonstrates a focus on the core section of the grain, which is a notable attribute of Arborio rice grain, renowned for its fullness and circular shape. The model's attention is directed towards the long and slender Basmati grains, which align with their uniquely elongated form. The visualization demonstrates that the model has a wide scope and considers the complete grain of Ipsala rice grain when classifying it, taking into account its overall form and size.

The heatmaps indicate a higher concentration in the central region of the Jasmine grains, which corresponds to the variety's shorter and rounder structure in comparison to Basmati. The Karacadag grains exhibit a noticeable emphasis on their ends and center, maybe indicating that the model uses these characteristics to distinguish this variety from others. The Grad-CAM results indicate that the model is proficiently recognizing and utilizing crucial morphological characteristics of the rice grains, which are necessary for precise categorization among the many types. The confusion matrix presented in Fig. 23 provides a concise summary of the ensemble model's performance, specifically when a dropout rate of 0.2 is applied. The majority of Arborio grains are accurately categorized as such, with 2201 true positives. However, there are a few instances where they are mistakenly labeled as Karacadag and Jasmine, with 15 and 13 false positives, respectively. The Basmati model has high accuracy in identifying the bulk of Basmati grains, with 1893 true positives. However, it also misclassifies a considerable proportion of grains as Ipsala, with 329 false positives. The model demonstrates outstanding classification accuracy, with a significant number of true positives (2344) and a very low number of false positives with just 1, identified as Arborio. Although there is some

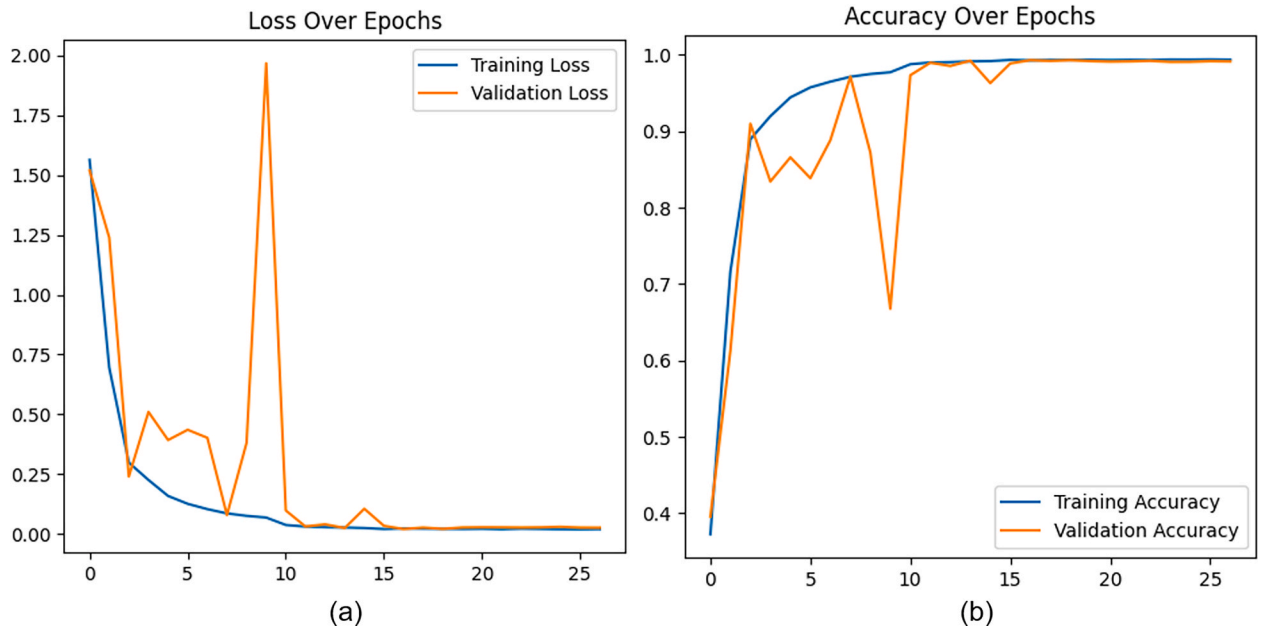


Fig. 20. Loss and accuracy plots of the ensemble model for dropout rate of 0.2.

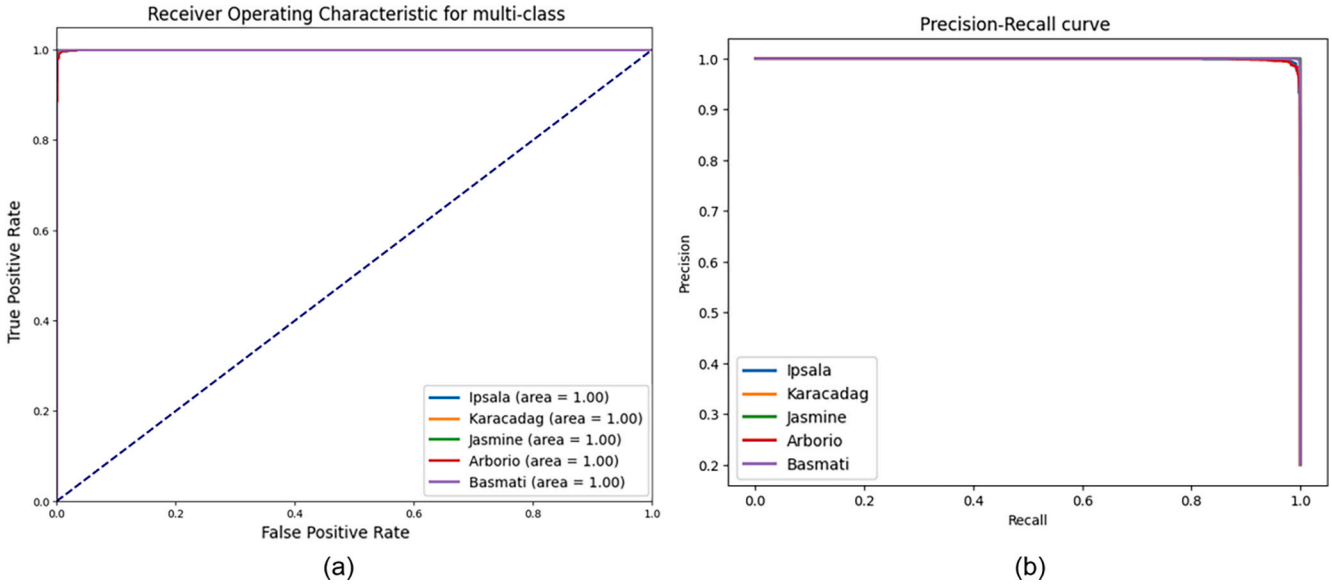


Fig. 21. ROC-AUC and precision-Recall curves of the ensemble model for the dropout rate of 0.2.

confusion with Karacadag (59 false positives), this variety nevertheless demonstrates a significant amount of accurate predictions (2192 true positives). The classification of Jasmine grains is mostly accurate, with 2136 true positives and only a few misclassifications. In general, the model exhibits a robust capacity to accurately categorize the rice grains, displaying a significant number of accurate positive predictions and relatively few incorrect positive predictions for each class label.

Table 10 presents a comparative analysis of classification using a dropout rate of 0.3, learning rate of 0.0001, and a batch size of 32. Precision improves across all rice types, with the most notable increase observed in Arborio, which rises to 99.24 %. Recall and sensitivity show varied trends: Ipsala maintains high recall at 99.64 %, while Basmati sees a slight decrease to 99.20 %. The F1-scores are strong for all classes, particularly Ipsala, which achieves 98.51 %. This suggests the model performs well across rice types, with Ipsala demonstrating strong generalization.

2.5.3. *Batch size*

Table 11 presents a comparative analysis of classification using a batch size of 64, dropout rate of 0.3, and learning rate of 0.001. The precision is exceptionally excellent for all classes, with Karacadag and Jasmine achieving flawless precision scores. The recall rate is likewise excellent, with Jasmine obtaining 99.96 %, demonstrating the model's superior ability to detect nearly all positive cases. The F1-scores exhibit robustness, indicating a well-balanced model, especially for Karacadag, Basmati and Jasmine. Generally, the table demonstrates that modifying the learning rate, implementing dropout, and optimizing the batch size can have a substantial impact on the model's performance, resulting in improved outcomes for particular classes.

2.6. *Comparison with built-in models*

In order to demonstrate the superior performance of the proposed

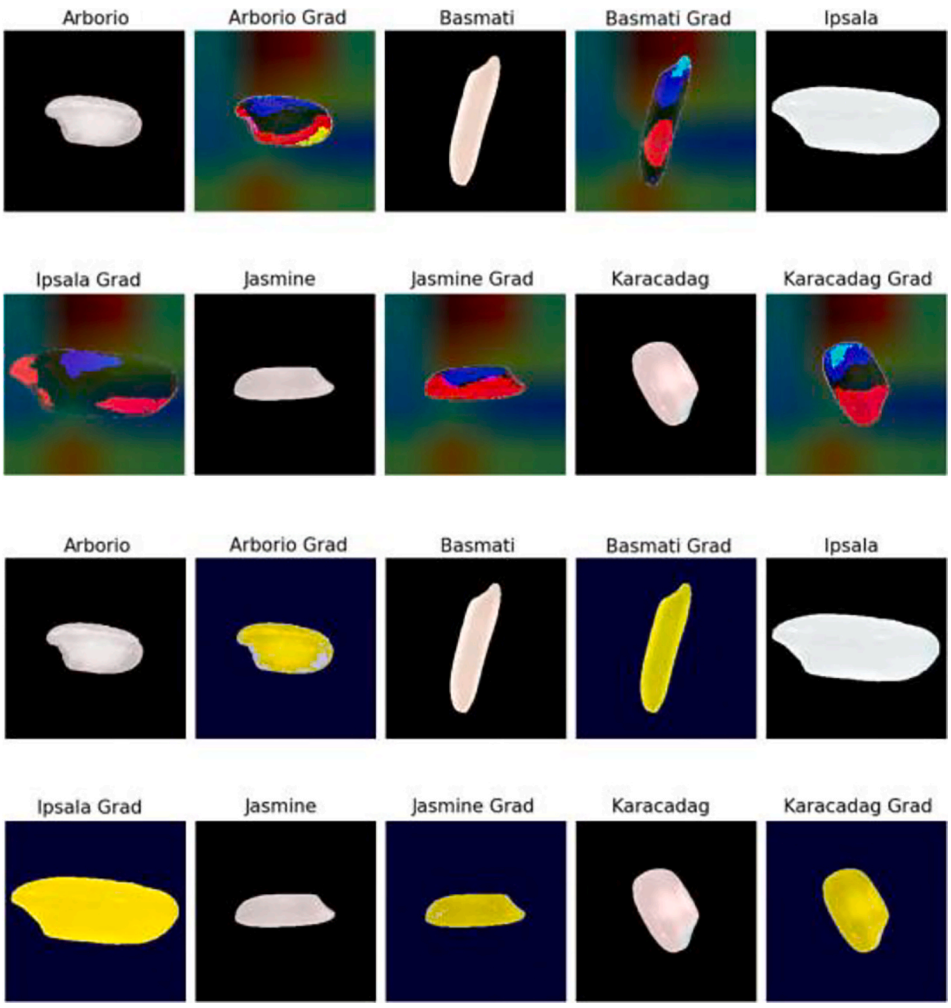


Fig. 22. Grad-CAM visualization of the ensemble model for dropout rate of 0.2.

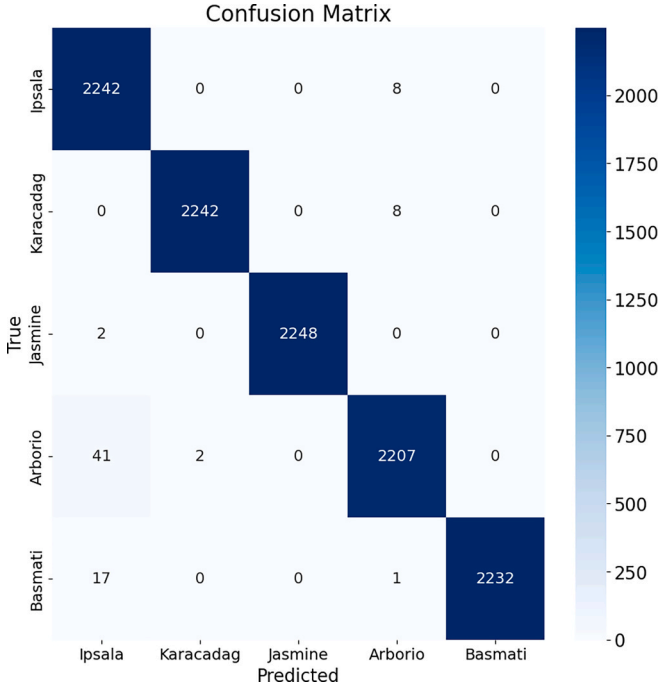


Fig. 23. Confusion matrix of the ensemble model for dropout rate of 0.2.

Table 10
Performance of the ensemble model across the different rice grains varieties using a dropout rate of 0.2.

	Ipsala	Karacadag	Jasmine	Arborio	Basmati
TP	2,242	2,242	2,248	2,207	2,232
FN	8	8	2	43	18
FP	60	2	0	17	0
TN	8,940	8,998	9,000	8,983	9,000
Precision	97.39 %	99.91 %	100.0 %	99.24 %	100.0 %
Recall	99.64 %	99.64 %	99.91 %	98.09 %	99.20 %
F1-score	98.51 %	99.78 %	99.96 %	98.66 %	99.60 %
Sensitivity	99.64 %	99.64 %	99.91 %	98.09 %	99.20 %
Specificity	99.33 %	99.98 %	100.00 %	99.81 %	100.0 %

Table 11
Performance of the ensemble model across the different rice grains varieties using a batch size of 64.

	Ipsala	Karacadag	Jasmine	Arborio	Basmati
TP	2,240	2,236	2,249	2,226	2,234
FN	10	14	1	24	16
FP	39	0	0	25	1
TN	8,961	9,000	9,000	8,975	8,999
Precision	98.29 %	100.0 %	100.00 %	98.89 %	99.96 %
Recall	99.56 %	99.38 %	99.96 %	98.93 %	99.29 %
F1-score	98.92 %	99.69 %	99.98 %	98.91 %	99.62 %
Sensitivity	99.56 %	99.38 %	99.96 %	98.93 %	99.29 %
Specificity	99.57 %	100.0 %	100.00 %	99.72 %	99.99 %

ensemble model, a comparative analysis with selected well-established pre-existing models is conducted. An analysis of such a comparison can uncover the advantages and drawbacks in several areas, offering valuable insights for future enhancements. Furthermore, in order to maintain fairness and precision in the comparison, the learning rate and data preprocessing are maintained at a consistent level when utilizing the pre-existing pre-trained models. Due to the limitation of image processing, which only allows input size of 50×50, the pre-trained models are deemed unsuitable. The ResNet50, VGG16, EfficientNetB0, and DenseNet201 models are chosen for this comparison alongside the proposed model in order to assess their performance in this experiment. These four pre-trained models possess distinctive architectural characteristics and are extensively employed in image recognition, classification, and other domains.

Fig. 24 depict the comparative performance of the different deep learning models in terms of accuracy and loss across training epochs. Fig. 24(a) denotes the performance comparison in terms of accuracy of training and validation. The proposed ensemble model exhibits a notable level of accuracy in both training and validation, with the curves of both sets of data closely aligning. This suggests that the model is able to generalize well without succumbing to overfitting. ResNet50, VGG16, and DenseNet likewise exhibit great accuracy. However, ResNet50 and VGG16 have discrepancies between training and validation accuracy, indicating potential overfitting. The accuracy of EfficientNet is comparatively lower, particularly during validation, suggesting either underfitting or a lack of suitability of the model for the given task. Fig. 24(b) denotes the performance comparison in terms of loss of training and validation. The proposed ensemble model exhibits low training and validation loss, indicating successful learning and strong generalization capabilities. The ResNet50, VGG16, and DenseNet models exhibit an early decline in loss, while ResNet50 and VGG16 display some fluctuations, suggesting possible instability in the learning process. The initial loss of EfficientNet is relatively significant, and although it reduces, it does not converge as tightly as other models, which is consistent with its lower observed accuracy. The proposed model's performance is clearly demonstrated by its exceptional precision and minimal loss, indicating its strong efficiency. It appears to accurately identify the fundamental patterns in the data without being influenced by random fluctuations or specific training data sets. By comparing the performance of the different models under the same training conditions, the proposed ensemble model performs well in this research study, as shown in Table 12.

Table 12 presents a comparative analysis of the selected pre-trained models based on specific parameters. The proposed ensemble demonstrates exceptional performance in all measures, with a minimal loss of 0.014, a high accuracy of 99 %, and an impressive F1 score of 99.77 %. ResNet50 demonstrates a moderate level of accuracy of 85 % and F1 score of 85.16 %, suggesting a reasonable level of performance, albeit not as high as the custom model. VGG16 demonstrates superior accuracy of 95 % and F1 score of 95.60 % compared to ResNet50, however it falls somewhat behind the proposed ensemble model. EfficientNetB0 exhibits a substantial decrease in all measures, characterized by a considerable loss value of 1.20, a low accuracy rate of 43 %, and a diminished F1 score of 37.07 %. DenseNet201 demonstrates strong performance with a high accuracy of 95.82 % and F1 score of 95.54 %, despite a relatively high loss of 10.89 %. The performance of each model is determined by a trade-off between precision and recall. The ensemble model has superior efficacy, whereas EfficientNetB0 encounters difficulties. Overall, the proposed ensemble model performed better in this study compared to the built-in models. While the built-in models provide a strong baseline, the proposed ensemble model happens to be better suited for this task.

While the proposed ensemble model outperforms the pre-trained models in terms of accuracy, loss, and F1 score, this improvement comes with notable trade-offs. The ensemble approach, which integrates features from multiple architectures, incurs higher computational complexity and longer training times compared to simpler models like VGG16 or ResNet50. The increased complexity of the ensemble model leads to longer training durations and greater resource usage, which can become a limiting factor in scenarios where computational resources are

Table 12
Fair comparisons between the proposed ensemble model and selected state-of-art pre-trained models.

Model	Loss	Accuracy	Precision	Recall	F1
Proposed model	0.014	99 %	99.76 %	99.56 %	99.77 %
ResNet50	0.34	85 %	85.16 %	85.19 %	85.16 %
VGG16	0.123	95 %	95.61 %	95.60 %	95.60 %
EfficientNetB0	1.20	43 %	52.72 %	43.57 %	37.07 %
DenseNet201	0.11	96 %	95.55 %	95.53 %	95.54 %

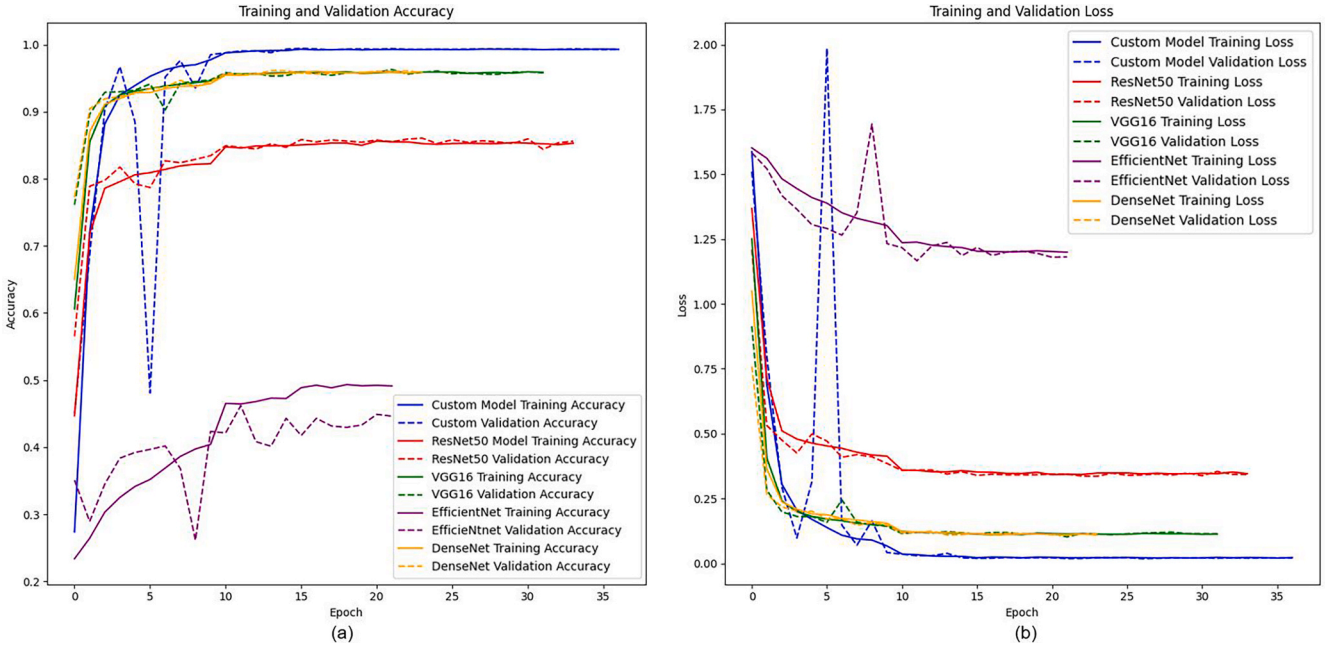


Fig. 24. Comparison analysis on Loss and Accuracy for the pre-trained model and proposed model.

constrained or real-time deployment is required. On the other hand, EfficientNetB0, though underperforming in this study, has lower complexity and a smaller model size, which may make it more suitable for resource-limited environments. Despite its lower accuracy, its reduced computational burden could make it a better choice in situations where speed or efficiency is prioritized over raw performance. These findings highlight the importance of balancing performance gains with resource efficiency, tailoring model selection to the specific requirements of deployment scenarios. In settings with limited computational resources or where real-time predictions are essential, lighter models such as EfficientNetB0 or ResNet50 might be preferred. However, for applications where higher accuracy is the primary concern, the proposed ensemble model proves to be the better choice, provided that computational resources allow for the added complexity.

2.7. Comparison with related literature

The performance of the proposed model is compared to published literature that has reported their findings in the field of rice grain variety classification. Table 13 provides a comprehensive comparison of these models, taking into account the dataset count utilized and the number of categories, and the F1-scores, which indicate the equilibrium between precision and recall.

Table 13 presents a comparison of F1 scores obtained from several studies on rice grain variety categorization using deep learning models. Chibhabha et al. (2022) obtained an F1 score of 97.92 % by employing an MVE classifier on a dataset consisting of 75,000 images distributed among 5 classes. Ilhan et al. (2021) employed a reduced dataset consisting of 3,810 data instances and 2 distinct categories. They achieved an F1 score of 93.04 % using a Deep Neural Network (DNN) classifier. Singh et al. (2023) utilized a dataset of equal size to the Chibhabha et al. (2022), however, achieved a lower F1 score of 86.00 % by employing a CNN. In the study conducted by Jeyaraj et al. (2022), they utilized a significantly smaller dataset consisting of 1,652 pieces and focused on 2 distinct classes. By employing the AlexNet model, they are able to attain an impressive F1 score of 86.00 %. Din et al. (2024) utilized a dataset of 4,748 instances and 5 distinct categories, achieved an F1 score of 98.20 % using a RiceNet classifier. This Study demonstrates the attainment of the highest F1 score, reaching an impressive 99.77 %, by employing a customized ensemble CNN model on a dataset including 75,000 images and 5 distinct classes. This study uses a rice grain dataset from the Kaggle repository, which has significant differences from other literature presented. These differences directly impact the model's performance metrics and comparison ability. The datasets differ in sample size, rice grain variety diversity, data quality, and phenotypic features, making it difficult to make direct comparisons. Publicly accessible data allows other researchers to replicate the result of this research and validate the findings, enhancing transparency and reproducibility. Despite these differences, the model's performance on a different dataset demonstrates its resilience and adaptability. The use of residual learning and multi-scale kernel architectures, along with bottleneck technique, improves feature extraction, lower computing costs, and model accuracy. The preprocessing technique used in the study is specifically customized for the Kaggle dataset. The model's architecture choice and hyperparameters have been optimized specifically for the Kaggle dataset. The performance on the Kaggle dataset demonstrates the proposed model's effectiveness and durability. Future research will involve evaluating the proposed model on additional datasets and conducting a comprehensive comparison analysis to further validate its benefits.

Table 13
Direct comparisons from reported literature and the proposed ensemble model.

Reference	Data Pieces	Class	Classifier	F1
Chibhabha et al. (2022)	75,000	5	MVE	97.92 %
Ilhan et al. (2021)	3,810	2	DNN	93.04 %
Singh et al. (2023)	75,000	5	CNN	86.00 %
Jeyaraj et al. (2022)	1,652	2	AlexNet	86.00 %
Din et al. (2024)	4,748	5	RiceNet	98.20 %
Proposed model	75,000	5	Ensemble CNN	99.77 %

3. Conclusion

This study explored the field of agricultural informatics with the aim of transforming the process of categorizing rice grains by utilizing sophisticated deep learning methods. Recognizing the crucial significance of rice as a staple meal for a large part of the world's population, particularly in the context of China's agricultural practices, the findings of the study have substantial implications for ensuring food security and promoting economic growth. The main focus of this study was to create a combined model that incorporates the features of residual learning and multi-scale kernel architectures. This model was specifically developed to leverage the unique advantages of hierarchical feature extraction and multi-scale feature analysis, representing a pioneering technique in this domain. This research distinguishes itself with the utilization of a vast dataset of 75,000 images, a customized attention mechanism, and a thorough evaluation of performance indicators. The ensemble model exhibited outstanding classification accuracy, offering the possibility of substantially diminishing the manual effort required for rice grain quality analysis. The effective utilization of Grad-CAM visuals also bolstered the model's reliability by providing an intelligible layer that delineates the key regions of interest crucial for categorization. Transparency is essential for using deep learning models in real-world scenarios, as stakeholders need to have trust in the automated procedures. Not standing for the significant advancements, the research identifies specific constraints. The dataset's exclusivity to Turkish rice grain varieties, albeit significant, restricts the model's applicability to other types of grains and varietals from diverse geographical regions. The high level of computational demand needed to evaluate a dataset of this size and execute intricate models is a challenge for scalability, especially in contexts with limited resources. In addition, although the ensemble model performed better than the prior models, it required significant hyperparameter tuning and iterative testing to obtain this level of performance. The applicability of this technique to other datasets or classification tasks without comparable rigorous optimization may be limited.

4. Limitation and future work

While the study presents significant advancements, it is important to recognize certain limitations. The dataset's specificity, albeit advantageous, restricts the model's applicability to other types of grains and varietals from diverse geographical regions. The dataset's exclusivity to Turkish rice grain varieties limits the model's ability to generalize across different rice grain varieties. Furthermore, the high level of computational demand required to evaluate a dataset of this size and execute intricate models presents scalability challenges, particularly in resource-limited settings. In addition, although the ensemble model performed better than prior models, it required substantial hyperparameter tuning and iterative testing to achieve its performance. The applicability of this technique to other datasets or classification tasks without similar rigorous optimization may yield less favorable results.

The study's constraints provide opportunities for future research avenues. Subsequent research should focus on expanding the dataset by including a wider range of rice grain cultivars from different regions across the globe. This will improve the model's resilience and suitability for worldwide applications. Additionally, sustainable agricultural practices, such as black biodegradable mulching could be explored to investigate their impact on rice quality and classification. Incorporating such practices into the dataset will not only improve the model's relevance but also align with global efforts to promote sustainable farming. Moreover, the computing requirements of the model, although feasible

in a research environment, may pose significant obstacles in practical scenarios, especially in developing nations. Efficiently improving the model to decrease computing burden while maintaining accuracy, continues to be a significant obstacle for upcoming versions.

In order to overcome these constraints and further the research, various options are suggested for future investigation. Inspired by the use of multi-modal feature extraction in fatigue recognition systems, Future work could integrate external datasets such as chemical composition and genetic markers into the classification process, a more comprehensive assessment of rice quality might be achieved. This would enable the development of more comprehensive models capable of processing multi-modal data for a holistic assessment of rice quality. Subsequent investigations should strive to augment the dataset by incorporating diverse rice grain cultivars originating from various global regions. By diversifying the dataset, the model's applicability and robustness will be enhanced, allowing it to be used with a wider variety of rice grain varieties. Incorporating the model into a categorization system that operates in real-time would be an extremely valuable progress. Implementing the model in actual sorting and grading technology has the potential to greatly simplify the process of rice quality control. By integrating multi-modal supplementary data, such as chemical composition and genetic markers, into the classification process, a more comprehensive assessment of rice quality might be achieved. It is important to investigate deep learning models capable of processing multi-modal data in order to combine these different sets of information. Further investigation can focus on improving the attention mechanism by incorporating additional data sources such as genetic markers or hyperspectral images. This hybrid approach could offer more precise classifications and broader applicability in real-world agricultural practices.

CRediT authorship contribution statement

Xudong Li: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Yutong Wang:** Writing – original draft, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Happy Nkanta Monday:** Writing – review & editing, Visualization, Validation, Supervision, Software, Investigation, Conceptualization. **Grace Ugochi Nneji:** Writing – review & editing, Visualization, Validation, Supervision, Software, Investigation.

Declaration of competing interest

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

Data availability

I have shared the link to the data in the manuscript.

References

Ahmed, T. Rahman, C. R. Abid, M., Mahmud, F., 2020. Rice grain disease identification using dual phase convolutional neural network-based system aimed at small dataset.

Aukkapinyo, K., Sawangwong, S., Pooyoi, P., Kusakunniran, W., 2019. Localization and classification of rice-grain images using region proposals-based convolutional neural network. Int. J.Autom. Comput. 1–14. https://doi.org/10.1007/s11633-019-1207-6.

Barbedo, J.G.A., 2016. A review on the main challenges in automatic plant disease identification based on visible range images. Biosyst. Eng. 144, 52–60. https://doi.org/10.1016/j.biosystemseng.2016.01.017.

Chen, J., Zhang, D., Zeb, A., Nanehkaran, Y.A., 2021. Identification of rice plant diseases using lightweight attention networks. Expert Syst. Appl. 169, 114514.

Chibhabha, K. Zvarevashe, L. K. Nyandoro, T. Matekenya and B. Mapako, "Classification of Rice varieties using DMLP-PCA inspired features with MVE Classifier," 2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT), Harare, Zimbabwe, 2022, pp. 1-7, doi: 10.1109/ZCICT55726.2022.10046040.

Cortes, C., Vapnik, V., 1995. Support-vector- networks. Machine Learn. 20 (3), 273–297. doi: 10.1007/BF00994018.

Dahl, G.E., Sainath, T.N., Hinton, G.E., 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 8609–8613. https://doi.org/10.1109/ICASSP.2013.6639346.

Din, N.M.U., Assad, A., Dar, R.A., Rasool, M., Sabha, S.U., Majeed, T., Islam, Z.U., Gulzar, W., Yaseen, A., 2024. RiceNet: a deep convolutional neural network approach for classification of rice varieties. Expert Syst. Appl. 235, 121214.

Ebrahimi, E., Mollazade, K., Babaei, S., 2014. Toward an automatic wheat purity measuring device: a machine vision-based neural networks-assisted imperialist competitive algorithm approach. Measurement 55, 196–205. https://doi.org/10.1016/j.measurement.2014.05.003.

Ilhan, U., Ilhan, A., Uyar, K., and Iseri, E. I., 2021. Classification of Osmancik and Cammeo Rice Varieties using Deep Neural Networks, In: 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2021, pp. 587–590. doi: 10.1109/ISMSIT52890.2021.9604606.

Jabeen, A., Subrahmanyam, D., Krishnaveni, D., 2023. The global lifeline: a staple crop sustaining two-thirds of the world's population. Agric. Arch.: Int. J. 2 (3), 15–18. https://doi.org/10.51470/AGRI.2023.2.3.15.

Jeyaraj, P.R., Asokan, S.P., Nadar, E.R.S., 2022. Computer-assisted real-time rice variety learning using deep learning network'. Rice Sci. 29 (5), 489–498. https://doi.org/10.1016/j.rsci.2022.02.003.

Jin, B., Zhang, C., Jia, L., Tang, Q., Gao, L., Zhao, G., Qi, H., 2022. Identification of rice seed varieties based on near-infrared hyperspectral imaging technology combined with deep learning. ACS Omega 7 (6), 4735–4749.

Kaya, E., Saritas, I., 2019. Towards a real-time sorting system: identification of vitreous durum wheat kernels using ANN based on their morphological, colour, wavelet and gaborlet features. Comput. Electron. Agric. 166, 105016. https://doi.org/10.1016/j.compag.2019.105016.

Koklu, M., Cinar, I., Taspinar, Y.S., 2021. Classification of rice varieties with deep learning methods. Comput. Electron. Agric. 187, 106285.

LaValley, M.P., 2008. Logistic regression. Circulation 117 (18), 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658.

Lin, N., Luo, X., Wen, J., Fu, J., Zhang, H., Siddique, K.H.M., Feng, H., Zhao, Y., 2024. Black biodegradable mulching increases grain yield and net return while decreasing carbon footprint in rain-fed conditions of the Loess Plateau. Field Crop Res 318, 109590. https://doi.org/10.1016/j.fcr.2024.109590.

Lin, P., Li, X., Chen, Y., He, Y., 2018. A deep convolutional neural network architecture for boosting image discrimination accuracy of rice species. Food Bioproc. Tech. 11 (4), 765–773. https://doi.org/10.1007/s11947-017-2050-9.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26. https://doi.org/10.1016/j.neucom.2016.12.038.

Liu, T., Wu, W., Chen, W., Sun, C., Chen, C., Wang, R., Zhu, X., Guo, W., 2016. A shadow-based method to calculate the percentage of filled rice grains. Biosyst. Eng. 150, 79–88. https://doi.org/10.1016/j.biosystemseng.2016.07.011.

Patrício, D.I., Rieder, R., 2018. Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. Comput. Electron. Agric. 153, 69–81. https://doi.org/10.1016/j.compag.2018.08.001.

Pan, H., Tong, S., Wei, X., Teng, B., 2024. Fatigue state recognition system for miners based on a multi-modal feature extraction and fusion framework. J. Latex Class Files, 2024. doi: 10.1109/TCDS.2024.3461713.

Rajalakshmi, R., Faizal, S., Sivasankaran, S., Geetha, R., 2024. RiceSeedNet: rice seed variety identification using deep neural network. J. Agric. Food Res. 16, 101062.

Sabanci, K., Kayabasi, A., Toktas, A., 2017. Computer vision-based method for classification of wheat grains using artificial neural network. J. Sci. Food Agric. 97 (8), 2588–2593. https://doi.org/10.1002/jsfa.8080.

Shrestha, B.L., Kang, Y.-M., Yu, D., Baik, O.-D., 2016. A two-camera machine vision approach to separating and identifying laboratory sprouted wheat kernels. Biosyst. Eng. 147, 265–273. https://doi.org/10.1016/j.biosystemseng.2016.04.008.

Singh, R., Sharma, N., and Gupta, R., 2023. Rice Type Classification using Proposed CNN Model. 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), 2023, pp. 1–6. doi: 10.1109/ViTECoN58111.2023.10157073.

Stephen, A., Punitha, A., Chandrasekar, A., 2023. Designing self attention-based ResNet architecture for rice leaf disease classification. Neural Comput. Applic. 35, 6737–6751. https://doi.org/10.1007/s00521-022-07793-2.

Sun, C., Liu, T., Ji, C., Jiang, M., Tian, T., Guo, D., Wang, L., Chen, Y., Liang, X., 2014. Evaluation and analysis the chalkiness of connected rice kernels based on image processing technology and support vector machine. J. Cereal Sci. 60 (2), 426–432. https://doi.org/10.1016/j.jcs.2014.04.009.

Tang, X., Zhao, W., Guo, J., Li, B., Liu, X., Wang, Y., Huang, F., 2023. Recognition of plasma-treated rice based on 3D Deep residual network with attention mechanism. Mathematics 11, 1686. https://doi.org/10.3390/math11071686.

Wang, Y., Wang, H., Peng, Z., 2021. Rice diseases detection and classification using attention based neural network and bayesian optimization. Expert Syst. Appl. 178, 114770.

Yang, X., Wang, J., Xia, X., Zhang, Z., He, J., Nong, B., Luo, T., Feng, R., Wu, Y., Pan, Y., Xiong, F., Zeng, Y., Chen, C., Guo, H., Xu, Z., Li, D., Deng, G., 2021. OsTTG1, a WD40 repeat gene, regulates anthocyanin biosynthesis in rice. Plant J. 107, 198–214. https://doi.org/10.1111/tpj.15285.

Zhang, D., Zhou, X., Zhang, J., Lan, Y., Xu, C., Liang, D., Wang, Z., 2018. Detection of rice sheath blight using an unmanned aerial system with high-resolution color and multispectral imaging. PLoS ONE 13 (5), e0187470. https://doi.org/10.1371/journal.pone.0187470.