



Hybrid convolutional neural network and multilayer perceptron vision transformer model for wheat species classification task: E-ResMLP⁺

Emrah Dönmez¹

Received: 9 November 2023 / Revised: 31 December 2023 / Accepted: 5 January 2024 / Published online: 20 February 2024
© The Author(s) 2024

Abstract

Wheat plant is one of the most basic food sources for the whole world. There are many species of wheat that differ according to the conditions of the region where they are grown. In this context, wheat species can exhibit different characteristics. Issues such as resistance to geographical conditions and productivity are at the forefront in this plant as in all other plants. The wheat species should be correctly distinguished for correct agricultural practice. In this study, a hybrid model based on the Vision Transformer (VT) approach and the Convolutional Neural Network (CNN) model was developed to classify wheat species. For this purpose, ResMLP architecture was modified and the EfficientNetV2b0 model was fine-tuned and improved. A hybrid transformer model has been developed by combining these two methods. As a result of the experiments, the overall accuracy performance has been determined as 98.33%. The potential power of the proposed method for computer-aided agricultural analysis systems is demonstrated.

Keywords CNN · Fine-tuning · Multi-layer perceptron · Vision transformer · Wheat species

Introduction

Wheat (*Triticum Aestivum*), as one of the basic food sources of humanity, is one of the most consumed grains in the world. The importance of wheat production has increased, especially due to the weakening of poor agricultural practices during the pandemic period and the subsequent Ukraine-Russia conflict. Both Ukraine and Russia are among the world's top 10 countries in wheat production [1]. Wheat has a great role in fighting hunger, especially in African countries with limited access to wheat. For these reasons, intensive incentives are applied for the cultivation of this plant, and productivity-increasing measures are taken. Wheat can have different types of seeds. According to the productivity conditions in the region where wheat will be planted, the appropriate one from these varieties is selected and planted. Wheat breeds can show different characteristics by adapting to the country or region where they are grown. The correct selection of wheat varieties is a critical issue in

plant breeding studies. Agricultural experts, farmers, etc. make manual selections to distinguish wheat types. Nowadays, it has become possible to make this selection with computer-aided systems [2, 3].

It is seen that computer-aided agriculture applications are becoming more widespread today [4–6]. The main purpose of these systems is mainly to increase productivity in agricultural products, to ensure product safety and to assist product breeding studies. To perform these tasks in computer-aided systems, decision support systems that model expert decisions are widely developed. In these systems, sequential decision mechanisms such as data acquisition, diagnosis and action/diagnosis are usually operated. The process of making certain inferences by taking and processing the image of the agricultural product is one of these mechanisms. Structures with certain characteristics are calculated on the product image and diagnostic processes related to its condition are modeled. Increasingly used artificial intelligence applications come to the fore to model these processes. Among artificial intelligence methods, deep learning approaches known as CNN on the image processing side and learning models such as transformers have recently been developed. CNN architectures are models that highlight features at each step through convolution layers on image data and process them through multiple processes until the final classification layer.

✉ Emrah Dönmez
emrahdonmez@bandirma.edu.tr

¹ Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bandırma Onyedi Eylül University, 10200 Bandırma, Balıkesir, Turkey

The most notable innovation in Transformers is the attention mechanism, which allows the model to focus on different parts of the input sequence when making predictions. This mechanism allows the model to effectively capture long-range dependencies.

There are significant studies on wheat plants in the literature. In these studies, researches have been carried out in a wide range from gene analysis of wheat seed, physical properties of the plant, morphological characteristics of the seed to its characteristic structure.

In wheat genus identification, the middle part of the seed is ground and gliadin proteins are extracted for electrophoretic analysis. In this analysis based on grain texture and hardness, phenol test is applied to extract the final genus characteristics [7]. Demyanchuk et al. [8] described the application of X-ray techniques for the non-destructive analysis of the internal structure of wheat grain by obtaining the morphometric characteristics of the kernels and applying a mathematical description of the "fusion technology" of the embryo and kernel. Güneş et al. [9] tried to identify the species and characteristics of wheat seeds grown in Türkiye by analyzing texture features with image processing techniques. Sabancı et al. [10] proposed a computer vision-based approach for the classification of bread and durum wheat grains using an artificial neural network method. In their study, they analyzed the physical characteristics of wheat (width, length, circumference), color (R, G and B channels) and texture features.

Martín-Gómez et al. [11] stated in their research that in shape analysis-based methods, the similarity of the analyzed images to a geometric shape is ignored. Based on this, they explained the J index approach, which is based on comparing the outlines of seed images with geometric shapes. With this index, they performed the classification task by calculating the percentage of similarity between a wheat seed image and the geometric shape used as a model. Zhou et al. [12] developed a deep learning-based method to recognize wheat kernel types using wheat images obtained with near-infrared (NIR) hyper-spectroscopy. In their study, they stated that they added a block for feature selection and an attention block that gives more importance to certain regions in the CNN model. Laabassi et al. [13] proposed an approach to recognize bread and durum wheat varieties from image data using five different pre-trained CNN models. For this purpose, they updated the parameters of CNN models by retraining all models.

Gao et al. [14] proposed a wheat variety recognition method using a convolutional neural network with image data from different growing periods of wheat before harvest. They stated that they improved the recognition accuracy by combining four CNN models in the study. Yasar [15] conducted a comparative analysis of CNN models used to recognize wheat varieties. For this purpose, image

data of five different wheat types were given as input to Inception-V3, Mobilenet-V2, and Resnet18 CNN models. As a result of the study, he stated that CNN models showed remarkable performance in wheat recognition. Zhao et al. [16] used the ensemble learning method to recognize the endosperm tissue on wheat seeds using images obtained with hyper-spectral imaging. In their study, they used spectral features as well as morphological features of wheat. They used CARS and SPA methods as feature extractors and support vector machine, nearest neighbor and decision tree methods for classification task.

When the studies in the literature are examined, it is seen that studies based on high-level feature analysis are also carried out in addition to approaches such as content and shape analysis. Similar to the studies based on seed analysis of wheat, there are also studies on breed recognition based on pre-harvest early growth period data. In some of these studies, data collected through conventional imaging systems, while in others, data collected through imaging systems such as near-infrared, hyper-spectral, etc. are analyzed. Some of the studies deal with two-class problems, while others deal with multi-class problems. On the other hand, it should be noted that the sample sizes analyzed also vary from study to study. In this study, a hybrid method developed with transformer and CNN approaches was used to analyze wheat image data obtained with conventional imaging systems.

In the tasks related to agricultural products, these methods shorten the decision-making processes considerably and facilitate timely intervention action to the necessary points by analyzing the data. The use of these methods for wheat, which is one of the important agricultural products, has significant potential. In this study, the potential use of the transformer method to extract the characteristics of wheat and classify them into certain classes is investigated and an improved transformer method is proposed. The contributions made within the scope of the study are as follows.

- The potential of using transformers with a dataset of wheat images was revealed.
- E-ResMLP⁺ architecture was developed as a transformer approach and classification of wheat data was performed.
- A hybrid model of ResMLP and EfficientNetV2b0 architectures was proposed and its performance on wheat data was analyzed.

The main motivation for this study was the fact that wheat is one of the main staple foods and production efficiency has reached a critical importance today. Especially the losses in production efficiency due to the pandemic and the subsequent conflict between Ukraine and Russia, which are among the top 10 producers of wheat production, have made access to wheat more difficult, which has increased

the importance of wheat. Due to increasing importance, it has become more necessary to develop smart systems that will directly contribute to increasing wheat productivity, ensuring product safety and breeding studies. On the other hand, a disruption in wheat production and a decrease in product efficiency have the potential to lead the world to famine.

The remaining part of the study is as follows. The second section describes the dataset used, the concepts of deep features and transformers, and the proposed methodology. In the next section, experiments are performed, and observations are noted. In the fourth section, a discussion is carried out based on the experimental results and other works in the literature. In the fifth and last section, general conclusions are given.

Material and method

Dataset

The dataset used in the study (Yasar, 2023) is a five-class dataset consisting of wheat seed images. The dataset contains a total of 8354 wheat seed images, of which "AYTEN ABLA" 1633, "BAYRAKTAR 2000" 1850, "HAMITBEY" 1624, "SANLI" 1600 and "TOSUNBEY" 1648 seed types. Table 1 below summarizes this situation about the dataset.

Deep features and transformers

Features on an image can be obtained in various ways depending on characteristics such as shape, color and texture. With these features, tasks such as recognizing, locating, or tracking an object searched on the image are performed. Instead of these attributes obtained with traditional approaches, nowadays, deep attributes with high representation level are obtained with CNN models, one of the deep learning networks. In the traditional approach, features are extracted by external methods and then sent to the classifier, whereas in deep learning networks, features are extracted automatically, and the classification layer is included in the network itself. Deep features are extracted as high-level

Table 1 Dataset summary

Species name	Quantity
AYTEN ABLA	1632
BAYRAKTAR 2000	1850
HAMITBEY	1624
SANLI	1600
TOSUNBEY	1648
Total	8354

features as a result of the operations performed in the convolutional layers of CNN models.

The EfficientNet CNN model [17] operates based on a scaling factor that optimizes model performance in response to increasing depth and width of the network. This CNN model is known as a widely used competitive model. In addition to the use of CNN models in image analysis, RNN (Recurrent neural network) based Transformers [18] methods are also used. Additionally, RNNs are specialized for handling sequential data, while ViTs are tailored for visual data using the transformer architecture. Transformer uses a framework characterized by layered self-attention and densely connected point layers in both encoder and decoder components. The structure of the model is shown in simple form in Fig. 1 below.

The encoder consists of 6 identical layers, where each layer consists of two subcomponents. The first subcomponent utilizes a multi-headed self-attention mechanism, while the next subcomponent contains a simple, location-specific, fully connected feed-forward network. The decoder likewise consists of a series of 6 identical layers. Each encoder layer has two subcomponents, while the decoder provides an additional third subcomponent. This third sub-layer transmits the multi-head attention to the output produced by the encoder stack. Let W^Q, W^V , and W^K be square matrices of size $T_{mdl} \times T_{mdl}$ representing the weights. These weights correspond to the vector set length for model T_{mdl} . The architecture with $Q(X) = X \times W^Q, V(X) = X \times W^V$ and $K(X) = X \times W^K$ is based on the attention function given below (1).

$$\text{Att}(Q, V, K) = \text{softmax}\left(\frac{Q(X)K(X)^T}{\sqrt{T_{mdl}}}\right) \times V(x) \tag{1}$$

In this equation modelled by the SoftMax function, the parameters Q, V and K represent a set of queries, values and keys, respectively. The scaling factor in the equation is expressed by $1/\sqrt{T_{mdl}}$. Multi-head attention allows the model to collectively focus on information from various representation subdomains and locations. Averaging inhibits this ability when applied with a single attention head. The width of the weight matrices of the individual heads is given by $T_{hd} = T_{mdl}/h$, where h is the number of attention heads. For a fixed parameter i (individual head), $Q_i(X) = X \times W_i^Q, V_i(X) = X \times W_i^V$ and $K_i(X) = X \times W_i^K$ and matrix dimensions $n \times T_{hd}$, the calculation for an individual head is done by Eq. (2).

$$\text{Att}_i(Q, V, K) = \text{softmax}\left(\frac{Q_i(X)K_i(X)^T}{\sqrt{T_{hd}}}\right) \times V_i(x) \tag{2}$$

The following equation structure (3) can be used to define the feed-forward layer.

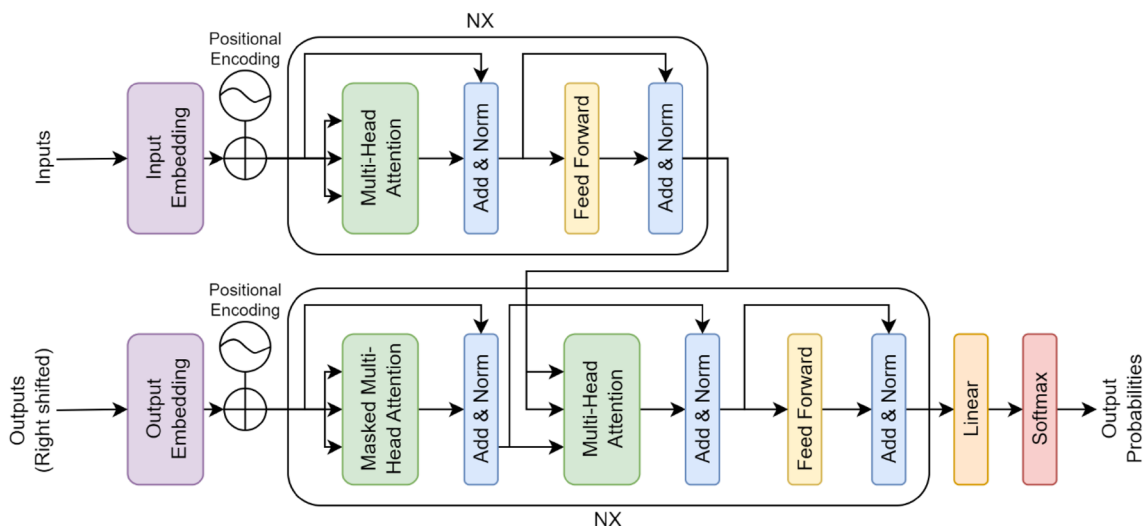


Fig. 1 Transformer architecture model

$$L_{FF} = ReLU(N_{ff}(X) \times A + K) \times B + L \tag{3}$$

In the above equation, N_{ff} is a two-vector normalisation layer with $a, b \in \mathbb{R}^{T_{mdl}}$. The parameters $A \rightarrow T_{mdl} \times T_{ff}$, $K \rightarrow n \times T_{ff}$, $B \rightarrow T_{ff} \times T_{mdl}$ and $L \rightarrow n \times T_{mdl}$ are weight matrices with given dimensions.

ResMLP (Residual Multi-Layer Perceptron) is a model inspired by vision transformer [19] approaches [20]. $N \times N$ non-overlapping patches are given as input to the model. Typically, the patch size is set to 16×16 . These patches are then passed to a linear layer. In this layer, a series of d-dimensional N^2 insertions are made. The N^2 embeddings are then input into a set of ResMLP units, resulting in a collection of d-dimensional output embeddings. Initially, a linear sublayer operates on patches, followed by a feed-forward sublayer acting on channels. The exclusion of the self-attention layers allows to replace the Layer Normalization with a simple Affine transformation (4), improving training stability.

$$Aff_{\alpha, \beta}(x) = Diag(\alpha)x + \beta \tag{4}$$

In this equation, α and β are learnable weight vectors. The Affine transformation process only adjusts the input by scaling and shifting individual elements. The main advantages of this approach are that there is no cost in inference time, and it does not rely on batch statistics.

EffV2b0-ResMLP⁺ hybrid method (E-ResMLP⁺)

In this study, the ResMLP architecture was modified, and a hybrid model was developed with the fine-tuned EfficientNetV2b0 model. In the EfficientNetV2b0 model,

the Batch Normalization layers were frozen during the fine-tuning phase. From the first epoch when the freezing process was cancelled, a decrease in accuracy was observed. In some cases, cancelling the freezing process in a part of the layers allows faster fine-tuning. In vision models, it is common practice to use Batch Normalization (BN) to normalize activations in the network [21] BN is typically applied on the non-normalized pre-activations X to generate the normalized pre-activations Y , (5). Following this, an affine transformation and a nonlinear ϕ are applied to obtain the post-activation Z , (6). In formal terms, this process is defined for each channel C .

$$Y_{...c} = \frac{X_{...c} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \tag{5}$$

$$Z_{...c} = \phi(\gamma_c Y_{...c} + \beta_c) \tag{6}$$

The parameter “•” in the equation represents an index placeholder. ϵ is the numerical stability constant of the batch normalisation. The values μ_c, σ_c are the mean and standard deviation for X in channel c , respectively. γ_c and β_c are the scale and shift parameters of the BN, respectively. There are two cases where the principles of BN are retained while eliminating the dependence on patch size. The first is when the BN operation in Eq. (5) is replaced by a group-independent normalisation method using Layer Normalisation (LN) or Group Normalisation (GN). Second, the activation operation in Eq. (6) is replaced by a proxy normalised activation approach. Considering all these cases, batch operations were empirically repeated in some layers during training.

Since each block in the architecture has a shortcut from the first layer to the last layer, the effect of the blocks was analyzed with an on/off approach. Blocks that had a negative impact on the final performance were disabled. After processing the image data through the finely tuned parameter weights in the whole network, the input to the multilayer sensors was generated.

The proposed multilayer perceptron processes a collection of N^2 d -dimensional input features arranged in a $d \times N^2$ matrix X and produces a set of N^2 d -dimensional output features arranged in a matrix Y using the following series of transformations (7) and (8).

$$Z = X + \text{Aff}\left(\left(\text{AAff}\left(A'\text{Aff}(X)^T\right)\right)^T\right)^T \tag{7}$$

$$Y = Z + \text{Aff}\left(\text{CGELU}\left(\text{BAff}\left(B'\text{Aff}(Z)\right)\right)\right) \tag{8}$$

In the above equations, A , A' , B , B' and C represent the main learnable weight matrices. The parameter matrix A has dimensions of $N^2 \times N^2$. This means that the "cross-patch" sublayer facilitates the exchange of information between patches. In addition, it indicates that the "cross-channel" feed-forward sublayer operates at each location. The intermediate activation matrix Z shares the same dimensions as the input and output matrices X and Y . Finally, the weight matrices B and C have the same dimensions as those of the Transformer layer, specifically $4d \times d$ and $d \times 4d$ respectively. The ResMLP+ architecture layer in the proposed hybrid model is given below, Fig. 2.

Experiments

Configuration

The computer hardware specifications used in the development and testing of the proposed method are as follows:

- Intel Xenon E5 Processor 2.2 GHz
- P4000 Quattro GPU
- 32 GB ECC Ram
- 512 GB SSD

In the proposed model, the patch size for the E-ResMLP+ module is set to 32×32 . The input images are 250×250 px in size and in RGB three-channel color space. The number of epochs for the experiments was set to 50. The data about 6683 (80%) were used for training in the total 8354 data in the dataset, 835 (10%) were used for validation and 836 (10%) were used for testing. Table 2 below shows the training, validation, and test data for each class in detail.

Table 2 Training, validation and test data, in detail

Species name	Train	Validation	Test	Quantity
AYTEN ABLA	1306	163	163	1632
BAYRAKTAR 2000	1480	185	185	1850
HAMITBEY	1299	162	163	1624
SANLI	1280	160	160	1600
TOSUNBEY	1318	165	165	1648
TOTAL	6683	835	836	8354

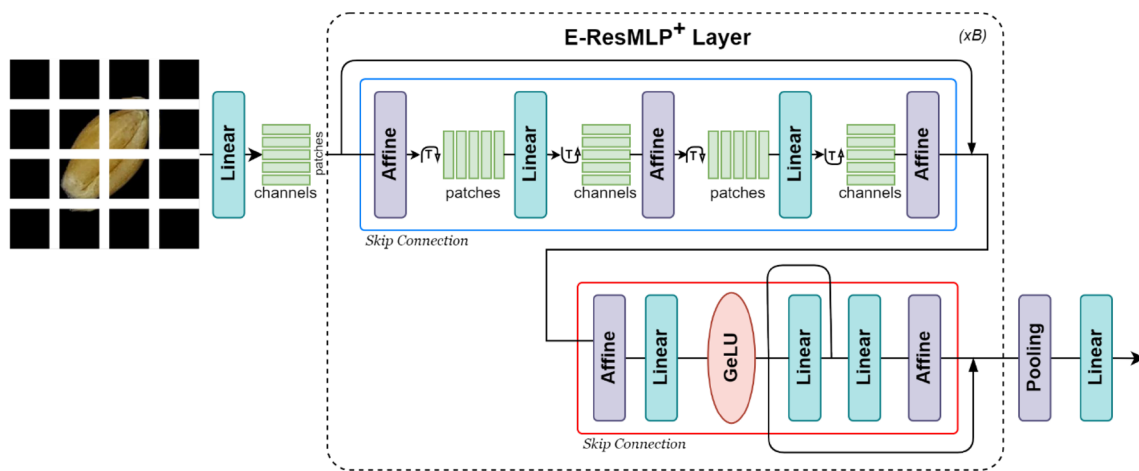


Fig. 2 Proposed E-ResMLP+ model

Performance metrics

Performance metrics correspond to the most basic parameters for evaluating the results of the experiments performed in this study. To evaluate the performance, we first used the confusion matrix containing the classification results of the test phase. A visualization of the confusion matrix is given in Fig. 3 below.

TP in this matrix means "True Positives" and refers to the correctly predicted positive class. FP stands for "False Positives" and refers to the incorrectly predicted positive class. TN stands for "True Negatives", which corresponds to the correctly predicted negative class. Finally, FN stands for "False Negatives" and corresponds to the incorrectly predicted negative class. Using all these values, the performance metrics Accuracy (9), Precision (10), Sensitivity (11) and F1-score (12) were used to evaluate each class separately. Balanced Accuracy (13), Misclassification Rate (14), Macro Average (15) and Weighted Average (16) metrics were used to evaluate overall performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{F1 - score} = \frac{2TP}{2TP + FP + FN} \quad (12)$$

Balanced Accuracy

$$= \frac{\text{Sensitivity} + \text{Specificity}}{2} \leftarrow \{\text{Specificity} = TN / (TN + FP)\} \quad (13)$$

$$\text{Misclassification Rate} = \frac{FN}{FN + TP} \quad (14)$$

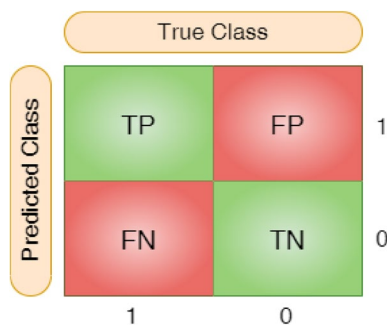


Fig. 3 Confusion matrix

$$\text{Macro Average} = \frac{\sum_{k=1}^c \text{Metric}_{(k)}}{c} \quad (15)$$

$$\text{Weighted Average} = \frac{N_{C_i} (\sum_{k=1}^c \text{Metric}_{(k)})}{TP + FP + TN + FN} \quad (16)$$

Accuracy is calculated as the total number of correct predictions divided by the total number of predictions and represents the performance in the relevant class. Precision is the number of correct predictions for the positive class divided by the total number of predictions for the positive class. It provides information about the prediction performance of the positive class. Sensitivity is the number of correct predictions for the positive class divided by the total number of positive class samples. Balanced Accuracy is an accuracy performance metric that expresses the overall performance of all classes. Misclassification Rate is the number of misclassified negative classes divided by the total number of positive classes. It provides information about misclassification. Macro Average is calculated using unweighted averages. In summary, it corresponds to the average value for each metric based on the number of classes. This value penalizes the model in case of inferior performance in other classes. Weighted Average is the weighted average. It is calculated by weighting the values of the metrics by the number of class instances and dividing by the total number of instances. It ensures that the class with the larger sample size is more effective than the average.

Experiment and results

The training and validation loss function graph of the hybrid model developed with the modified ResMLP and fine-tuned EfficientNetV2b0 models within the scope of the experiments is presented in Fig. 4 below. It can be stated that the loss function value tends to decrease during training and validation. While the amount of decrease is sharper at the beginning of training and validation, the loss function change has become more stable in training since the fifth epoch. On the other hand, it is observed that the validation loss function change takes values in more variable ranges. Training time is determined as approximately 266 s.

The training and validation performance graph resulting from the proposed model is shown in Fig. 5 below. When the graph is analyzed, the training and validation performance tends to increase. While a sudden increase is observed in the beginning, similar to the loss function pattern, it can be stated that the change becomes more stable in training and more variable in verification.

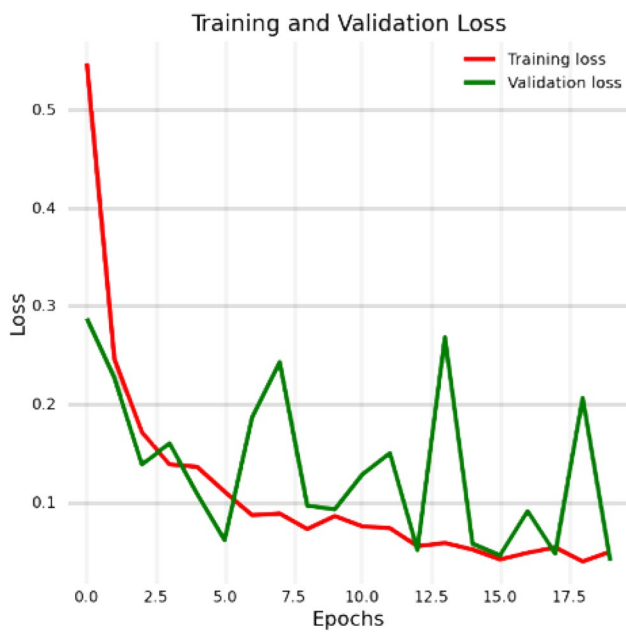


Fig. 4 Loss value variation for training and validation processes

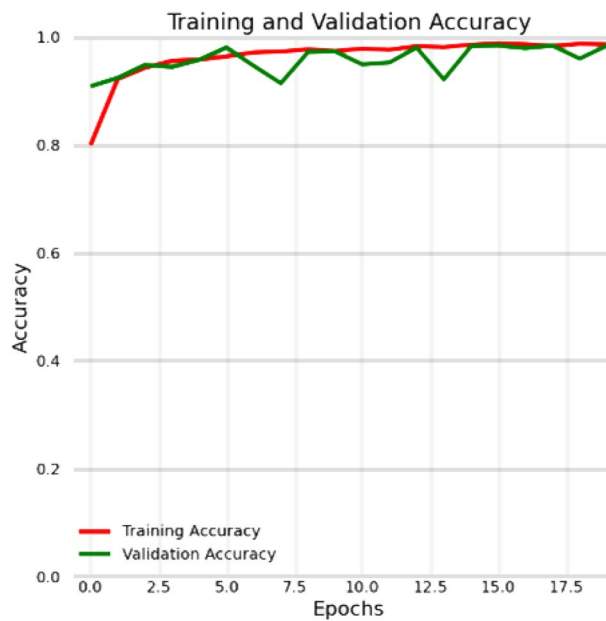


Fig. 5 Performance value change for training and validation

Another performance indicator of the model, the ROC change graph of the training and validation process is given in Fig. 6 below. The ROC change gives us a clue about the change between the true positive rate (TPR) and the false positive rate (FPR). In this case, it is understood that there are changes between TPR and FPR very close to the value of 1 during all epochs in both training and validation.

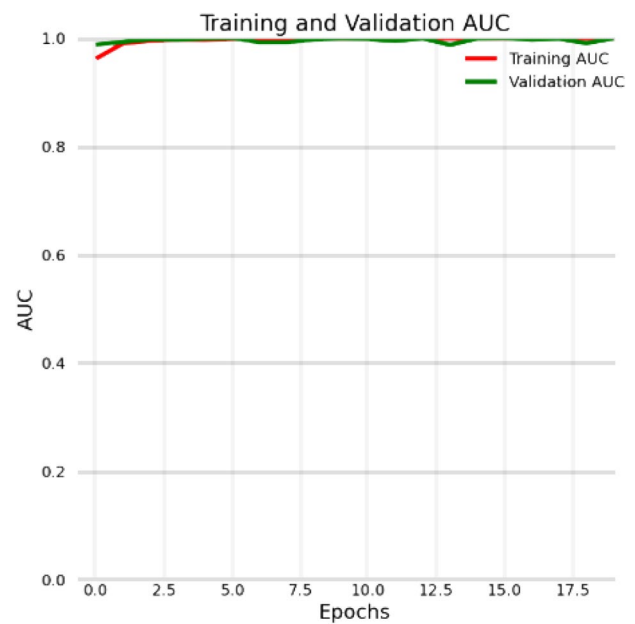


Fig. 6 ROC value change for training and validation

The graph of the confusion matrix obtained using the test data is shown in Fig. 7 below. According to the confusion matrix, it can be said that the proposed method provides high discrimination among almost all classes. Only in the TOSUNBEY class, 7 samples classified as HAMITBEY are noteworthy. Even in this case, the accuracy value is at a remarkable level of 95.15%. According to the table, it is understood that 3 samples were misclassified in the AYTEN ABLA class, 1 each in BAYRAKTAR 2000, HAMITBEY and SANLI classes and 8 samples in the TOSUNBEY class.

In Table 3 below, the performance metric values obtained within the scope of the experiments are presented separately for each class. When the table is analyzed, it can be seen that sensitivity and accuracy values are equal. This situation arises because the discrimination performance of the relevant class in multi-class problems is equal to the sensitivity value. When the table is examined, it is determined that accuracy values above 98% are obtained in AYTEN ABLA class, 99% in BAYRAKTAR 2000, HAMITBEY and SANLI classes and 95% in TOSUNBEY class. It is seen that the precision values in AYTEN ABLA and SANLI classes are perfectly obtained with 1.

The macro average and weighted average values obtained within the scope of the experiments are presented in Table 4 below. In all macro and weighted average values, all performance metrics reach values above 98%.

The balanced accuracy value, which shows the overall performance value, was calculated as 0.9833 (98.33%) using the confusion matrix. Thus, the overall misclassification rate is calculated as 0.0167 (1.67%). These values

Fig. 7 Confusion matrix of the proposed model

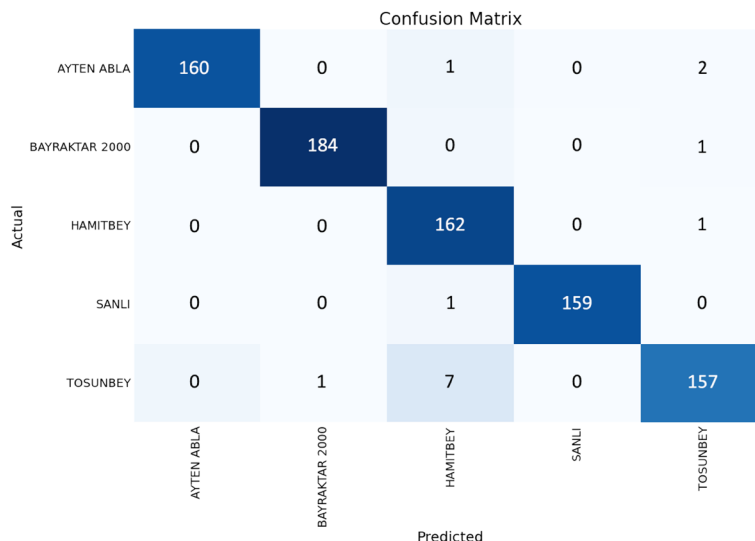


Table 3 Performance metrics for classes

Species	Accuracy	Precision	Sensitivity	F1-score	Miss. Rate
AYTEN ABLA	0.9816	1.0000	0.9816	0.9907	0.0184
BAYRAKTAR 2000	0.9946	0.9946	0.9946	0.9946	0.0054
HAMITBEY	0.9939	0.9474	0.9939	0.9701	0.0061
SANLI	0.9938	1.0000	0.9938	0.9969	0.0062
TOSUNBEY	0.9515	0.9752	0.9515	0.9632	0.0485

Table 4 Average performance metric values

P. Metrics	Precision	Sensitivity	F1-score
Macro Avg	0.9834	0.9831	0.9831
Weighted Avg	0.9836	0.9833	0.9833

represent the final classification performance when all classes are considered.

To observe the performance of the proposed model more effectively, an additional experiment was conducted using the dataset used by Laabassi et al. [13]. There are four classes in total in the data used in the experiment, two of which are hard wheat (Simento and Vitron) and two are soft wheat (ARZ and HD). As a result of the test, the values of the confusion matrix were formed as shown in Table 5 below.

As a result of the experiment, the balanced accuracy value was around 98.29%. Performance metrics obtained using the confusion matrix are also given in Table 6 below.

After the general experiments, experiments were also carried out with different CNN models to compare the overall performance of the proposed method. Multi-class balanced accuracy values obtained using these models are given in the Table 7 below.

Table 5 Confusion matrix metrics for additional dataset

Species (Output/Target)	SIMENTO	FP	TN	FN	Total
SIMENTO	967	10	3	4	984 (98.27%)
VITRON	4	666	1	1	672 (99.11%)
ARZ	1	2	415	6	424 (97.88%)
HD	5	3	14	1059	1081 (97.96%)
Total	977 (98.98%)	681 (97.80%)	433 (95.84%)	1070 (98.97%)	3161 (98.29%)

Table 6 Performance metrics for classes (4 classes)

Species	Accuracy	Precision	Sensitivity	F1-score	Miss. Rate
SIMENTO	0.9898	0.9827	0.9898	0.9862	0.0102
VITRON	0.9780	0.9911	0.9780	0.9845	0.0220
ARZ	0.9584	0.9788	0.9584	0.9685	0.0416
HD	0.9897	0.9796	0.9897	0.9847	0.0103

Table 7 Balanced accuracy values: CNN models vs Proposed model

Models	Balanced accuracy
InceptionV2	0.9369
ResNet-101	0.9455
EfficientNetb0	0.9474
This Study	0.9833

Discussion and observations

Wheat (*Triticum aestivum* L.) is the most important food source among the major food sources for the whole world. Wheat species have different characteristics in terms of resistance to conditions (seasonal, etc.) and productivity. In this respect, it is very important to distinguish the species to carry out correct planting and breeding studies. When the studies on wheat in the literature are examined, it is seen that studies have been carried out on different methods, equipment and data sets. In Table 8 below, the prominent studies in the literature and this study are presented comparatively.

When the studies in the table are examined, species identification was performed according to the signal frequency obtained from impedance measurement hardware [22]. Species identification was performed by extracting deep features of wheat images obtained with hyper-spectroscopy imaging over Near Infrared (NIR) images [12]. On the other hand, a significant number of species identification studies have been carried out using transfer learning approaches on wheat images obtained with standard camera [13, 15, 23]. In this study, the ResMLP architecture, which is a visual transformer approach, was modified and used in a hybrid way with the EfficientNetV2b0 CNN model. The developed method is called E-ResMLP+. In the proposed work, no

data preprocessing, transfer learning, feature selection or fusion and no special hardware were used. In the experiments performed with the developed method, the overall performance value was obtained as 98.33%.

Conclusion

As a staple food source, wheat is a plant species for which species discrimination is important. In this study, the ResMLP architecture was modified by utilizing the power of the visual transformer method to perform species discrimination. The modified architecture was used as a hybrid with the EfficientNetV2b0 CNN model. As a result of the experiments, the overall classification performance was calculated as 98.33%. With the proposed approach, significant levels have been achieved in performance metrics. The study demonstrates the potential power of the transformer approach for computer-aided agricultural analysis systems. It should be noted that performance improvement is also promising as a result of improving the architecture of transformer methods and the hyperparameters of CNN models.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability The data set used within the scope of the study will be shared upon request.

Declarations

Conflict of interest The authors declare that there are no known conflict of interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 8 Comparison of featured and proposed studies

Study	Method	Hardware	Data	Performance
Zhou et al. (2020)	CNN feature selection CNN with Attention Support vector machine	Near Infrared (NIR) hyper-spectroscopy imaging	147,096 (30 classes)	93.01%
Laabassi et al. (2021)	Transfer Learning Pre-trained models	Standard camera	31,606 (4 classes)	95.68%
Yasar (2023)	Image preprocessing Transfer Learning Pre-trained models	Standard camera	8354 (5 classes)	97.67%
Kılıcarslan and Kılıcarslan (2023)	Image preprocessing GLCM + Color + Deep Features Transfer Learning + Feature selection Machine Learning	Standard camera	8354 (5 classes)	98.65%
This Study	Visual Transformer Modified ResMLP (E-ResMLP+) Fine-Tuned EfficientNetV2b0	Standard camera	8354 (5 classes)	98.33%

Compliance with ethics requirements No ethical compliance is required within the scope of this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Production of wheat worldwide 2022/2023. In: Statista. <https://www.statista.com/statistics/267268/production-of-wheat-worldwide-since-1990/>. Accessed 11 Sep 2023
2. Sabanci K, Aslan MF, Ropelewska E et al (2022) A novel convolutional-recurrent hybrid network for sunn pest-damaged wheat grain detection. *Food Anal Methods* 15:1748–1760. <https://doi.org/10.1007/s12161-022-02251-0>
3. Unlarsen MF, Sonmez ME, Aslan MF et al (2022) CNN–SVM hybrid model for varietal classification of wheat based on bulk samples. *Eur Food Res Technol* 248:2043–2052. <https://doi.org/10.1007/s00217-022-04029-4>
4. Diker A, Elen A, Közkurt C et al (2023) An effective feature extraction method for olive peacock eye leaf disease classification. *Eur Food Res Technol*. <https://doi.org/10.1007/s00217-023-04386-8>
5. Dönmez E (2022) Enhancing classification capacity of CNN models with deep feature selection and fusion: a case study on maize seed classification. *Data Knowl Eng* 141:102075. <https://doi.org/10.1016/j.datak.2022.102075>
6. Dönmez E, Kılıçarslan S, Közkurt C et al (2023) Identification of haploid and diploid maize seeds using hybrid transformer model. *Multimedia Syst*. <https://doi.org/10.1007/s00530-023-01174-y>
7. Wrigley CW (1976) Single-seed identification of wheat varieties: Use of grain hardness testing, electrophoretic analysis and a rapid test paper for phenol reaction. *J Sci Food Agric* 27:429–432. <https://doi.org/10.1002/jsfa.2740270507>
8. Demyanchuk AM, Grundas S, P.Velikanov L, et al (2013) Identification of Wheat Morphotype and Variety Based on XRay Images of Kernels. In: *Advances in Agrophysical Research*. IntechOpen
9. Güneş EO, Aygün S, Kırıcı M, et al (2014) Determination of the varieties and characteristics of wheat seeds grown in Turkey using image processing techniques. In: 2014 The Third International Conference on Agro-Geoinformatics. pp 1–4
10. Sabanci K, Kayabasi A, Toktas A (2017) Computer vision-based method for classification of wheat grains using artificial neural network. *J Sci Food Agric* 97:2588–2593. <https://doi.org/10.1002/jsfa.8080>
11. Martín-Gómez JJ, Rewicz A, Goriewa-Duba K et al (2019) Morphological description and classification of wheat kernels based on geometric models. *Agronomy* 9:399. <https://doi.org/10.3390/agronomy9070399>
12. Zhou L, Zhang C, Taha MF et al (2020) Wheat Kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method. *Front Plant Sci*. <https://doi.org/10.3389/fpls.2020.575810>
13. Laabassi K, Belarbi MA, Mahmoudi S et al (2021) Wheat varieties identification based on a deep learning approach. *J Saudi Soc Agric Sci* 20:281–289. <https://doi.org/10.1016/j.jssas.2021.02.008>
14. Gao J, Liu C, Han J et al (2021) Identification method of wheat cultivars by using a convolutional neural network combined with images of multiple growth periods of wheat. *Symmetry* 13:2012. <https://doi.org/10.3390/sym13112012>
15. Yasar A (2023) Benchmarking analysis of CNN models for bread wheat varieties. *Eur Food Res Technol* 249:749–758. <https://doi.org/10.1007/s00217-022-04172-y>
16. Zhao W, Zhao X, Luo B et al (2023) Identification of wheat seed endosperm texture using hyperspectral imaging combined with an ensemble learning model. *J Food Compos Anal* 121:105398. <https://doi.org/10.1016/j.jfca.2023.105398>
17. Tan M, Le QV (2020) EfficientNet: rethinking model scaling for convolutional neural networks
18. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
19. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16 x 16 words: transformers for image recognition at scale
20. Touvron H, Bojanowski P, Caron M et al (2021) ResMLP: feedforward networks for image classification with data-efficient training. *IEEE Trans Pattern Anal Mach Intellig*. <https://doi.org/10.1109/TPAMI.2022.3206148>
21. Masters D, Labatie A, Eaton-Rosen Z, Luschi C (2021) Making efficientnet more efficient: exploring batch-independent normalization, group convolutions and reduced resolution training
22. Kandala CVK, Govindarajan KN, Puppala N et al (2014) Identification of wheat varieties with a parallel-plate capacitance sensor using Fisher's linear discriminant analysis. *J Sensors*. <https://doi.org/10.1155/2014/691898>
23. Kılıçarslan S, Kılıçarslan S (2023) A comparative study of bread wheat varieties identification on feature extraction, feature selection and machine learning algorithms. *Eur Food Res Technol*. <https://doi.org/10.1007/s00217-023-04372-0>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.