European Food Research and Technology
https://doi.org/10.1007/s00217-024-04537-5

ORIGINAL PAPER

Performance of vision transformer and swin transformer models for lemon quality classification in fruit juice factories

Sezer Dümen[1] · Esra Kavalcı Yılmaz[2] · Kemal Adem[2] · Erdinç Avaroglu[3]

Received: 6 January 2024 / Revised: 12 March 2024 / Accepted: 16 March 2024

Abstract

Assessing the quality of agricultural products holds vital significance in enhancing production efficiency and market viability. The adoption of artificial intelligence (AI) has notably surged for this purpose, employing deep learning and machine learning techniques to process and classify agricultural product images, adhering to defined standards. This study focuses on the lemon dataset, encompassing 'good' and 'bad' quality classes, initiate by augmenting data through rescaling, random zoom, flip, and rotation methods. Subsequently, employing eight diverse deep learning approaches and two transformer methods for classification, the study culminated in the ViT method achieving an unprecedented 99.84% accuracy, 99.95% recall, and 99.66% precision, marking the highest accuracy documented. These findings strongly advocate for the efficacy of the ViT method in successfully classifying lemon quality, spotlighting its potential impact on agricultural quality assessment.

Keywords Lemon quality · Deep learning · Vision transformer · Swin transformer

Introduction

Determining the classification of agricultural products according to their physical characteristics, such as shape, size, color, and quality status, is a common process both nationally and internationally. Lemon is an important agricultural product that requires appropriate classification due to its annual production cycle and nutritional value as a rich vitamin C and antioxidant source. Lemon production in Turkey has been steadily increasing over the years and has been experiencing the fastest growth in recent years. According to the data obtained, approximately 750,550 tons of lemons were produced in 2015, 1,188,517 tons in 2020, and 1,550,000 tons in 2021. With 1,188,517 tons of lemon production in 2020, Turkey accounted for 41.1% of European lemon production. Of this lemon production, 45% was used for domestic consumption and 54.8% for exports. Accurate grading of lemons and early detection of diseases is very important, as disease and poor quality cause a decline in the plant market [1]. Traditional manual methods of classification and detection are not only slower, more laborious, and less efficient, but can also be easily affected by external factors such as fatigue, experience, and the psychological state of the experts. This can lead to misclassification and detection, which can reduce the market value of the product. To overcome these challenges, artificial intelligence and computer vision technologies are now being used to improve the accuracy of classifications, reduce erroneous processes, and increase efficiency in operations [2]. This approach also allows experts to focus on other areas of their expertise, leading to increased economic prosperity for the country.

Machine learning and deep learning have made significant progress due to the availability of large amounts of data. These methods have become important in numerous research areas, such as image processing and data classification. As a result, machine learning and deep learning methods have been applied in various fields, including medicine, industrial

✉ Kemal Adem
kemaladem@sivas.edu.tr

Sezer Dümen
sdumen@sivas.edu.tr

Esra Kavalcı Yılmaz
esra.kavalci@sivas.edu.tr

Erdinç Avaroglu
eavaroglu@mersin.edu.tr

[1] Department of Electrical and Electronics Engineering, Sivas University of Science and Technology, Sivas, Turkey

[2] Department of Computer Engineering, Sivas University of Science and Technology, Sivas, Turkey

[3] Department of Computer Engineering, Mersin University, Mersin, Turkey

Published online: 15 April 2024

applications, energy systems, and agriculture [3–9]. Farmers can use machine learning and deep learning-based applications to monitor crop production processes in natural and greenhouse environments. These applications are widely used in areas such as plant quality diagnostics, soil analysis, insect diagnostics, disease detection, and treatment to improve agricultural productivity [10].

Classification of agricultural products is an important issue, especially in understanding the impact of irregularly shaped products on consumer preferences and wastage issues. In this context, artificial intelligence-based studies using the shape and quality of products offer significant potential in this field. In previous studies, statistical methods such as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) have been used to classify product quality using agricultural product images such as carrots [11] and apricots [12]. When these classification processes are analyzed in general, it is seen that the QDA method is more successful.

In addition, fast and accurate classification processes such as disease detection, quality, and variety determination in agricultural products are among the factors that will contribute to increasing agricultural productivity. In this context, many studies have been conducted to classify the quality and type of products using artificial intelligence methods. In some of these studies, datasets consisting directly of product images were used [13, 14], while in some studies, classification processes were performed using statistical features obtained using product images [15–20]. These image and numerical data sets are used with various machine learning and deep learning algorithms to classify products quickly and accurately.

There are many classification studies in the literature using lemon images, as we have used in this study. These studies use various deep learning and machine learning methods to perform successful and fast classification processes [21, 22]. In classification processes, whether the input image is grayscale or color is among the factors that will affect the success of the process [23]. In addition to the images of lemons, the detection of diseases from the leaf images of the lemon plant in the agricultural field before the harvest is collected is one of the important factors for early treatment/spraying [24]. With the classification of leaf images taken from the field, more efficient agricultural activities can be realized with lower costs. In deep learning methods, the number of data in the dataset also contributes to successful results. For this reason, synthetic data can be obtained using data augmentation methods such as Generative Adversarial Networks (GANs) to obtain more successful results in case of insufficient data. By increasing the number of data with these methods, data scarcity problems can be solved. Thus, more successful results can be obtained in studies on small data sets [25].

Motivation and our model

The quality of the products is evaluated by size, color, shape, presence of disease or rot. In this case, the external appearance of the products can be considered as the main factor affecting the market. Therefore, the correct classification of products is of great importance. Performing the classification processes manually by experts with computer vision methods will not only increase product efficiency and market value, but also ensure better quality management of experts' working time. In addition, since the work of experts is more easily affected by external factors, the use of computer systems will increase the success rate of classification processes. In this study, various transformer methods and deep learning methods were used to classify lemon images. Akerlof's theory of the lemon market assumes that buyers in second-hand markets do not have sufficient information about the quality of the products they want to buy. In this case, sellers can sell lower quality products at the same price as high-quality products. Therefore, buyers will not be willing to pay a higher price for high-quality products, which can lead to a market crash [26]. This study also reveals the quality difference of products through lemon quality detection and therefore makes an indirect contribution to this problem.

In the study, the number of data was increased by performing rescaling, random zooming, random flipping, and random rotation operations on the dataset before training. Afterward, transformer methods such as Vision Transformer (ViT), Swin Transformer and deep learning methods such as Xception, ResNet-50, InceptionV3, NASNet-Mobile, EfficientNet-B5, InceptionResNetV2, ResNet-152, DenseNet-201 were applied to the obtained dataset. As a result of the study, it was observed that more successful results were obtained with transformer methods.

Novelties and contributions

The novelties of our study are as follows:

- The study offers an alternative viewpoint on how to apply the ViT architecture to images acquired in the field of agriculture.
- To increase the success of the study, the dataset was augmented using four different methods: rescaling, random zoom, random flip, and random rotation.
- For the first time in this study, transformer methods were used for classification on the lemon dataset.
- Within the scope of the study, classification processes were performed with deep learning methods and trans-

former methods, and it was observed that higher success was achieved when transformer methods were used.
– The success value obtained with the ViT method used in the study is 99.84%, which is the highest accuracy value in the literature.

The rest of the paper, Section ''Material and methods'', Material and methods, provides information about the dataset and methods used in the study. Section ''Results and discussion'', Results and discussion, contains all the analysis and discussion of the results obtained from the study. The full summary of the study is given in Section ''Conclusion'', Conclusion.

Material and methods

The Lemon Quality Dataset [27] is utilized in this study to classify the quality of lemons. The dataset comprises 2076 images, with each image being 300 × 300 pixels in size. The images depict lemons of varying sizes and qualities, with 951 images classified as bad and 1125 as good quality.

Deep learning methods

In the study, eight different deep learning models were used. The focus of the study is on powerful deep learning models such as Xception, ResNet-50, InceptionV3, NASNetMobile, EfficientNet-B5, InceptionResNetV2, ResNet-152, and DenseNet-201. These models are particularly favored for their ability to effectively reduce gradient loss, their parallel processing capacity, their lightweight model structure, their low computational power requirements, and their ability to easily adapt to different datasets. Increasing the interaction between layers through dense connections is another important advantage that improves the overall performance of the selected models. This feature highlights the potential to achieve more effective results by optimizing knowledge transfer in the learning process. Considering these main advantages, this study aims to achieve more effective and powerful results in the field of image classification using the deep learning models mentioned in the study. In this section, the two models with the highest accuracy, EfficientNet-B5 and DenseNet-201, are mentioned.

EfficentNet-B5

EfficientNet is one of the most effective artificial neural network models developed by Google and used in the field of deep learning. There are seven versions in total, and the most important feature between versions is the number of layers used. The model is determined according to the size of the input image [28].

EfficientNet-B5 is the medium-sized version of the EfficientNet family, and they are especially used in image classification operations. EfficientNet-B5 achieves higher accuracy than the smaller EfficientNet-B4 but requires less computing power than the larger EfficientNet-B6. EfficientNet-B5 performs particularly well when working with high-resolution images [29].

EfficientNet balances CNN characteristics such as width, depth, and resolution with a technique called composite scaling. The EfficientNet family offers models in various sizes, and each model is optimized for hardware with a specific computing power [30]. The architecture of the EfficientNet-B5 model used in the study is presented in Fig. 1 [31].

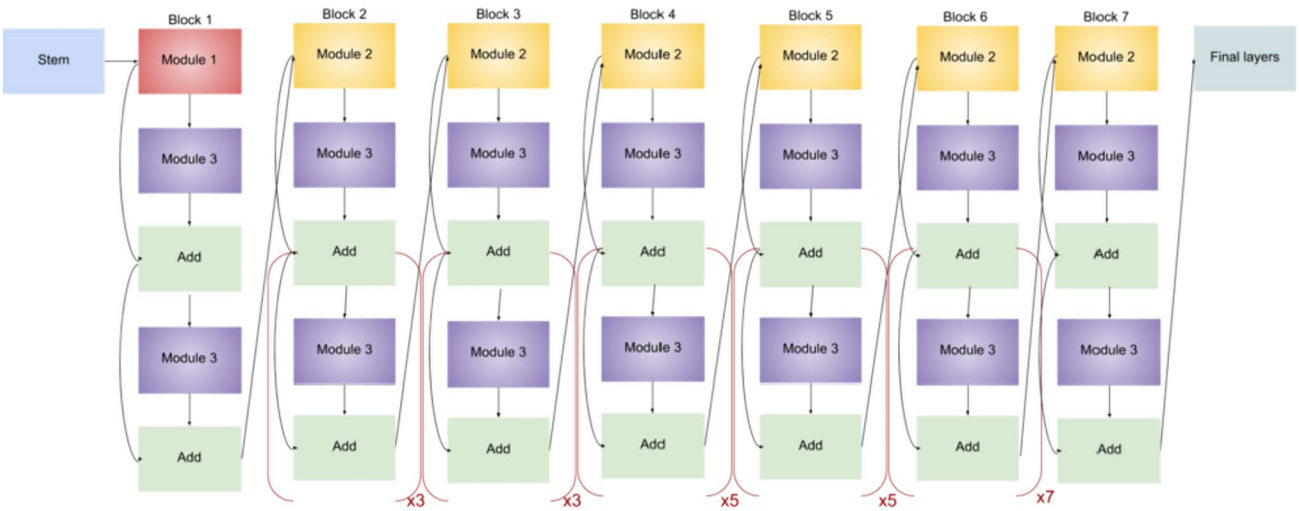

Fig. 1 Architecture of EfficientNet-B5

DenseNet-201

DenseNet is a convolutional neural network with direct feed-forward connections from each layer to all other layers. Its main advantage is that it minimizes the lost gradient resolution and overfitting in deep inspection problems that can occur with small training datasets [32]. In this architecture, the output from each layer receives all the outputs from the previous layer and processes it together with a unique "growth rate" parameter. Thus, the number of inputs used in each layer increases and better performance can be achieved using fewer parameters [33]. DenseNet-201 a member of the DenseNet Family with 201 layers, uses a condensed network to build models that are easy to train and highly efficient. It shows high performance due to the fact that the current layer can directly access the feature maps of all previous layers [34]. The architectural model of the DenseNet-201 method used in this study is shown in Fig. 2 [32].

Transformer methods

In this study, transformer methods, which are Vision Transformer and Swin Transformer, are used. In contrast to conventional deep learning architectures, the Vision Transformer [35] and Swin Transformer [36] models represent novel advancements. The Swin Transformer model, in particular, presents a recent innovation achieved through the integration of a shifted-window mechanism into the Vision Transformer framework. Both architectures employ attention mechanisms to efficiently manage computational resources by focusing solely on pertinent regions within the image. Tailored for image processing endeavors, these models excel in image classification tasks by mitigating information loss within the convolutional layers of convolutional neural network architectures. Consequently, the adoption of these innovative models was deemed imperative for the present study. The explanations of these methods are given in this section.

Vision Transformer (ViT)

Vision Transformer (ViT), introduced by Dosovitskiy et al. [35], is a deep learning model used in the field of machine learning and computer vision. ViT uses the transformer architecture, replacing the previously widely used convolutional neural network (CNN) architecture, and uses a set of attention mechanisms to process images. ViT processes images by making them into "patches". This reduces the size of images and makes it possible to process higher dimensional images compared to previous models. It also uses attention mechanisms instead of CNNs in learning the features of images. In this way, it can provide training with higher accuracy and less data [37].

ViT is a way to extend the traditional transformer application to image categorization without including any data-specific design to generalize non-textual modalities. ViT models a series of image segments into a semantic tag using transformer's encoder module for classification. While traditional CNN designs often use filters with a limited receptive field, ViT's attention mechanism focuses on different parts of the image and interprets information throughout the entire image. Thanks to these features, ViT is the first image recognition model to beat traditional CNN designs (e.g., limited filter usage). ViT architecture consists of Embedding Layer,

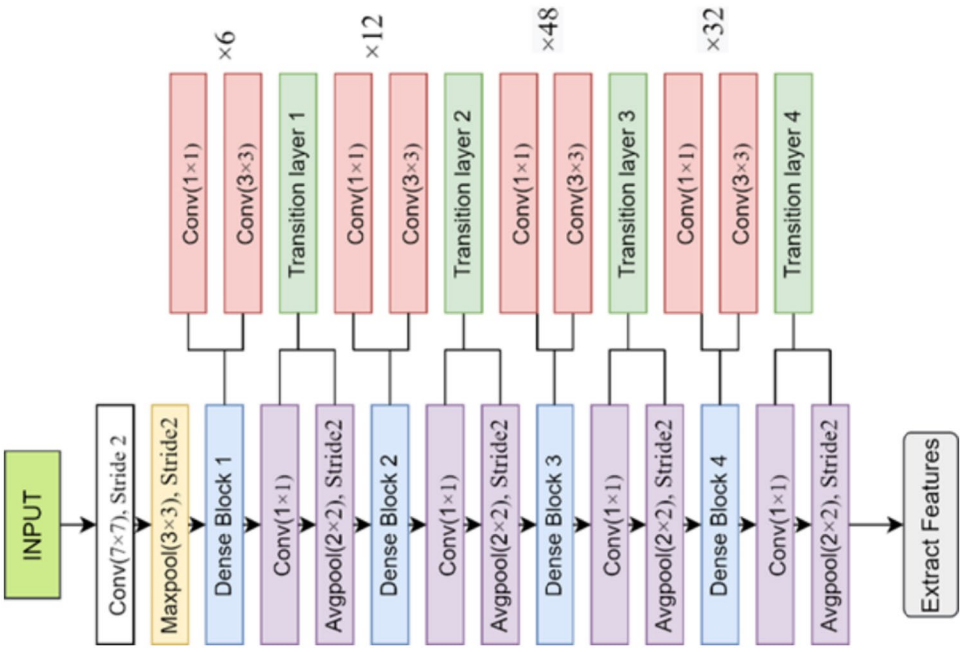

Fig. 2 Architecture of DenseNet-201

Encoder, and Final classifier head layers [38]. The Vit architecture model is presented in Fig. 3 [39].

Swin transformer

Swin Transformer is a method developed to facilitate the learning of large-scale image classification models [36]. It is used for various purposes such as region-level object detection, pixel-level semantic segmentation, and image-level image classification [40]. Swin Transformer introduces a novel methodology aimed at mitigating the memory constraints often encountered in prior image classification models. The core of this approach lies in the segmentation of image blocks into distinct block groups, followed by the implementation of a sophisticated "shift mechanism" designed to facilitate the seamless exchange of information among blocks within each group. By strategically orchestrating these inter-block interactions, Swin Transformer effectively minimizes memory overhead while concurrently enhancing scalability, thereby surpassing the limitations observed in conventional large-scale image classification architectures. This innovative paradigm not only optimizes resource utilization but also empowers the model to handle increasingly complex datasets with unprecedented efficiency and efficacy. Moreover, the shift mechanism employed by Swin Transformer exhibits a remarkable adaptability to varying spatial relationships within the image, enabling it to capture intricate patterns and dependencies across different regions. This dynamic information exchange fosters a more comprehensive understanding of the image content, leading to improved classification accuracy and robustness against distortions. By harnessing the power of block-wise processing coupled with strategic information flow, Swin Transformer redefines the paradigm for large-scale image classification, setting new benchmarks in both memory efficiency and performance scalability. Furthermore, the modular nature of Swin Transformer facilitates seamless integration with existing architectures and allows for straightforward customization to suit specific application domains. The simple architecture of Swin Transformer is given in Fig. 2 [41] (Fig. 4).

Results and discussion

In the study, a dataset consisting of a total of 2076 images, 1125 good quality, and 951 bad quality was used to determine the lemon quality. Before the training with deep learning and transformer methods, data augmentation is applied to images. These methods are rescaling, random

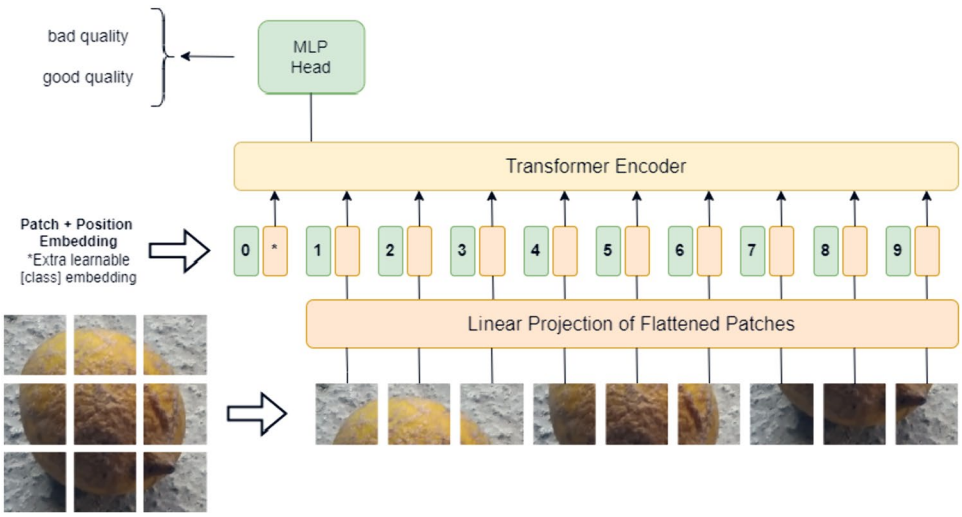

Fig. 3 Architecture of Vision Transformer

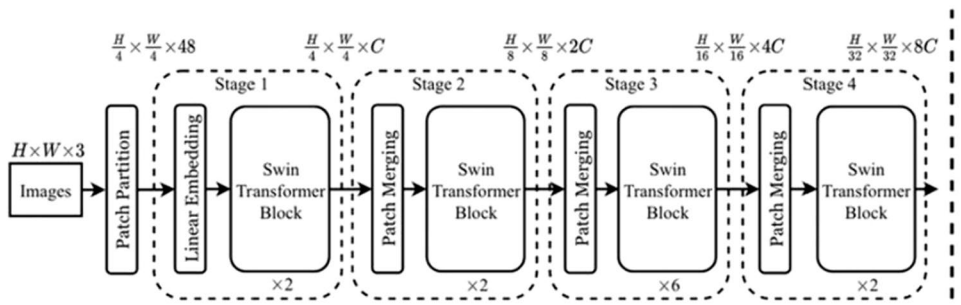

Fig. 4 Architecture of Swin Transformer

zoom, random flip, and random rotation. To determine the quality of lemons transformer methods which are Vision Transformer (ViT), Swin Transformer, and deep learning methods, which are Xception, ResNet-50, InceptionV3, NASNetMobile, EfficientNet-B5, InceptionResNetV2, ResNet-152, DenseNet-201 methods are used. For the performance evaluation of machine learning and deep learning models, the data augmented dataset is divided into 70% training (5812 images) and 30% testing (623 images). Of the 5812 lemon images used in the training phase, 3150 images were classified as good-quality lemons and 2662 images were classified as bad-quality lemons. Of the 623 lemon images used in the testing phase, 338 images were classified as good-quality lemons and 285 images were classified as bad-quality lemons. The block diagram of our proposed model including data augmentation and deep learning methods is shown in Fig. 5.

The hybrid models implemented in Fig. 5 were tested using Python on a computer with an i9 12,950 processor, RTX 3080TI graphics card, and 32 GB RAM.

Evaluation criteria

In the field of machine learning, evaluating model performance is crucial for assessing the effectiveness and generalization capabilities of trained models. There are several metrics for performance evaluation for classification methods which are validation accuracy, validation loss, precision, recall, and F1 score [42]. In this study, validation accuracy, validation loss, precision, and recall metrics were used to evaluate the performance of deep learning methods. These metrics provide valuable insights into the model's ability to make accurate predictions on unseen data and are widely employed in model selection and performance comparison.

Validation loss refers to the measurement of the discrepancy between the model's predicted output and the true target values on a validation dataset, which consists of examples that were not used during the model training phase. The validation loss is typically computed using a specific loss function that quantifies the dissimilarity between predicted and true values [43]. By monitoring the validation loss, researchers and practitioners can gauge the model's ability to generalize to unseen data and detect signs of overfitting. A low validation loss indicates that the model is performing well on the validation set, implying that it is effectively capturing the underlying patterns and regularities in the data. A high validation loss, on the other hand, suggests that the model may be struggling to generalize or is overfitting to the training data [44]. The goal is to minimize the validation loss, as it reflects the model's performance on unseen instances and serves as a proxy for its performance in real-world scenarios. Loss calculation is given in Eq. 1.

$$Loss = \frac{1}{N}\sum_{i}^{N} f\left(\hat{y}_i, y_i\right) \tag{1}$$

Loss functions are mathematical expressions that measure the difference between the predicted values of a model and the actual values and try to minimize this difference. In this function given in Eq. 1;

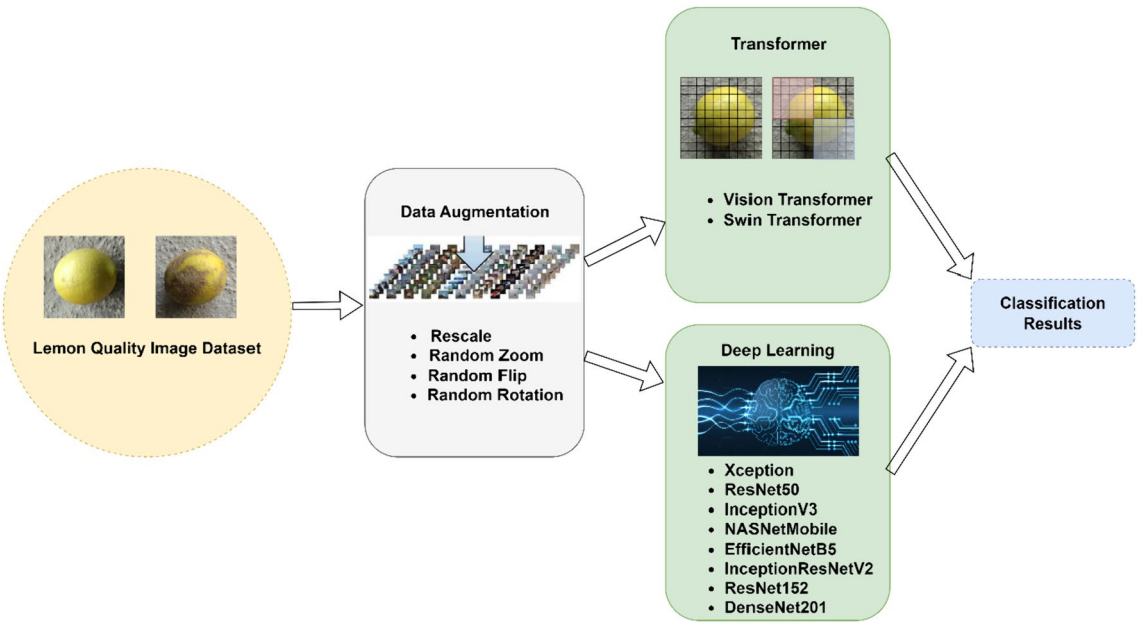

Fig. 5 The block diagram of our proposed model

- N is the total number of data points,
- i refers to each sample in the data series,
- $\hat{y}_i$ is the predicted value of the model,
- y_i refers to the actual value,
- $f(\hat{y}_i, y_i)$ is an error function that measures the difference between, $\hat{y}_i$,y_i.

In this formula, the error measure, denoted by the error function $f(\hat{y}_i, y_i)$, is calculated for each data point separately. Then these errors are averaged for all data points to obtain an overall loss value. This overall loss value is used to evaluate the performance of the model and to optimize it during model training.

Accuracy, on the other hand, measures the proportion of correctly predicted instances from the total number of examples in the dataset. It is a metric that is particularly relevant in classification tasks, where the model's output is a class label or a probability distribution over classes [44]. The accuracy provides a measure of how well the model can classify unseen data, offering insights into its overall predictive capabilities. A high accuracy implies that the model is making accurate predictions on the validation set, correctly assigning instances to their respective classes. Conversely, a low validation accuracy suggests that the model may struggle with generalization or encounter difficulties in distinguishing between different classes. Similar to loss, the objective is to maximize the accuracy, indicating that the model is performing well on unseen data [45]. Accuracy calculation is given in Eq. 2.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} x100 \tag{2}$$

$$Recall = \frac{TP}{TP + FN} x100 \tag{3}$$

$$Precision = \frac{TP}{TP + FP} x100 \tag{4}$$

Included in the equations are *TP* true positive, *TN* true negative, *FP* false positive, and *FN* false negative. Loss and accuracy are complementary metrics that provide a comprehensive evaluation of a trained model's performance. While loss quantifies the model's prediction errors in a continuous manner, accuracy provides a more interpretable measure of classification correctness. Precision is the proportion of correctly predicted positive instances (true positives) out of the total instances predicted as positive. It measures the accuracy of positive predictions, indicating how reliable the model is when it identifies positive samples. Recall, also known as sensitivity or true positive rate, is the proportion of correctly predicted positive instances (true positives) out of all actual positive instances. It measures the model's ability to identify all positive samples, indicating how effectively it captures the relevant instances. These metrics play a vital role in model evaluation, enabling researchers and practitioners to compare different models, assess their generalization capabilities, and make informed decisions about model selection and hyperparameter tuning [46]. After applying data augmentation techniques to the dataset consisting of lemon images, eight different deep learning models, namely Xception, ResNet-50, InceptionV3, NASNetMobile, EfficientNet-B5, InceptionResNetV2, ResNet-152, DenseNet-201, were applied. The training parameters used while applying these models are given in Table 1.

Table 1 Hyperparameters of deep learning models

Hyperparameters	Value
Epoch	20
Learning rate	0.01
Batch size	8
Input shape	300×300
Optimizer	Adam
Dropout	0.1
Activation function	ReLU
Output function	Softmax

Table 2 The results of the experimental studies carried out in the study

Models	Epoch	Recall	Precision	Accuracy (%)	Loss
Xception	20	98.35	97.81	98.02	0.0736
ResNet-50		96.53	95.96	96.17	0.1123
InceptionV3		98.02	97.59	97.83	0.0822
NASNetMobile		96.95	96.38	96.62	0.1086
EfficientNet-B5		**99.29**	**98.86**	**99.03**	**0.0382**
Inception-ResNetV2		98.82	98.19	98.43	0.0694
ResNet-152		97.21	96.66	96.84	0.0985
DenseNet-201		**98.96**	**98.49**	**98.65**	**0.0563**

Highest accuracy values are in bold

As a result of experimental tests, the values with high classification accuracy and low loss value were chosen as the training parameters for the deep learning models shown in Table 1. The value "20" was selected for the *Epoch*, which displays how many times deep learning models have been trained using the training dataset. A value of "0.01" was chosen for the *learning rate*, which affects the learning capacity and the learning time. The *batch size* value used to update the weights at each training step was found to be "8" when computing the loss function. The *Optimizer* function "Adam" was selected to enhance the weights. The *Dropout* value, which breaks the connection between neurons, was set at "0.1" to avoid overfitting. The outcomes of the deep

learning models are provided in Table 2 as a result of the training settings chosen. The best accuracy and lowest loss values that each deep learning model was able to achieve after training are shown in this table.

As seen in Table 2, as a result of the application of eight different deep learning models to lemon images, it is seen that the EfficientNet-B5 and DenseNet-201 models have higher accuracy values than the other deep learning models given in the table. In addition to the accuracy value, recall and precision values are also calculated. Considering these values, it has been observed that the recall value is higher than the accuracy value, and the precision value is lower than the accuracy value. The recall value gives the proportion of correctly classified positive samples. Because false negative classifications can cause serious problems, the recall value is very important in classification processes. False negative classifications can overlook what the object is and create obstacles to making the right decision. Precision value is an important evaluation metric used in classification processes where false positive classification is a priority. The fact that the recall value is higher than the accuracy value in this study indicates that good-quality lemons are classified with high accuracy. Considering the products used in fruit juice factories, it is thought that it is meaningful that the recall value is higher than the accuracy value in our study since it is used in fruit juice production in medium-quality products. In addition to deep learning models, recently popular Vision Transformer models have also been applied to lemon images to increase the accuracy values. Training parameters used when applying transformer models are given in Table 3.

In the Vision Transformer, we train the model by splitting the image into patches. *Patch size* refers to the size of these patches. *Projection dimension* refers to the length of the vector that we project these separated patches with the linear projection method. After the projection, the vectors we have obtained are placed in the multi-head attention layers in the transformer encoders, and it is decided how much attention should be paid to the result, considering how much it affects the result. *The number of heads* parameter refers to the number of heads in the multi-head attention layers. A transformer layer includes normalization, multi-head attention, and MLP layers. The *number of transformer layer* parameters indicates the number of these transformer layers. After the transformers comes the MLP layers, and the *MLP units* parameter refers to the size of these MLP layers. Swin Transformer has a shifted-window structure compared to Vision Transformer. This shifted-window mechanism processes the image by selecting windows on the image that we have divided into patches and shifting these windows. The *window size* parameter expresses the size of the windows on these patches. *Shift size* refers to how many pixels these windows will be shifted. The *label smoothing* parameter in Swin Transformer refers to a correction factor that is used to smooth the sharp target distribution, usually caused by the hard coding of the labels. This factor usually takes a value in the range [0, 1]. 0 means that label smoothing is not applied, while a value of 1 means maximum label smoothing [36]. These parameters used in Vision Transformer and Swin Transformer method are obtained by grid search method. For each parameter, various parameter spaces were searched and the best combination of parameters was found as shown in Table 3. The evaluation metrics obtained as a result of applying the transformer models and the two most successful deep learning models with the parameters specified in Table 3 to the dataset consisting of lemon images are given in Table 4.

As seen in Table 5, it is seen that transformer models are more successful than deep learning models. Among the transformer models, it is seen that the Vision Transformer model performs a more successful classification than the Swin Transformer model with an accuracy value of 99.84%. To show the consistency of the accuracy and loss values of these four models, box plot graphs are drawn and shown in Figs. 6 and 7.

Figures 6 and 7 show the average loss and accuracy values for the dataset prepared to determine lemon quality. Experimental evaluations were carried out on EfficientNet-B5,

Table 3 Hyperparameters of transformer models

Hyperparameters	Vision Transformer	Swin Transformer
Epoch	100	100
Learning rate	0.0001	0.0001
Batch size	8	8
Optimizer	Adam	Adam
Input shape	300×300	300×300
Patch size	15	10
Projection dimension	225	200
MLP units	1800,900	1024
Number of transformer layers	5	–
Number of heads	45	8
Window size	–	5
Shift size	–	1
Label smoothing	–	0.1
Activation function	ReLU	ReLU
Output function	Softmax	Softmax

Table 4 The results of the experimental studies carried out in the study

	Recall	Precision	Accuracy (%)	Loss
EfficientNet-B5	99.29	98.86	99.03	0.0382
DenseNet-201	98.96	98.49	98.65	0.0563
Vision Transformer	**99.95**	**99.66**	**99.84**	**0.0070**
Swin Transformer	99.38	99.12	99.23	0.0174

Highest accuracy values are in bold

Table 5 Comparison of the studies with Lemon Quality Dataset on literature

Authors	Year	Microarray data	Number of data	Method	Accuracy (%)
He et al	2021	Lemon	1847	VGG16	95.44
Pramanik et al	2021	Lemon	314	Xception	94.34
Hernandez et al	2021	Lemon	913	CNN	92
Bird et al	2022	Lemon	2690	VGG16	83.77
			2690 + 400	VGG16 after CGAN	88.75
Sharma et al	2022	Lemon	3000	CNN + LSTM	94.2
Yılmaz et al	2023	Lemon	2076	SAE- CNN	98.96
Proposed Method	2023	Lemon	2076	VisionTransformer	99.84

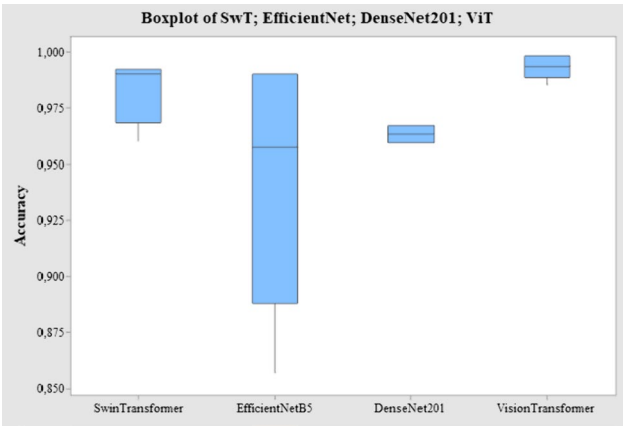

Fig. 6 Box plot graph of the accuracy values of the models

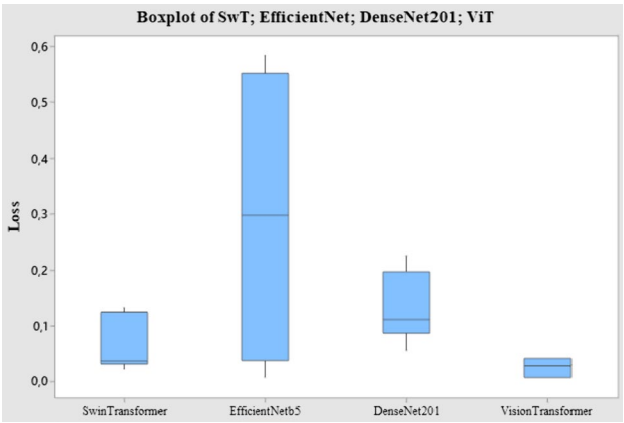

Fig. 7 Box plot graph of the loss values of the models

DenseNet-201 deep learning architectures and Vision Transformer, Swin Transformer transformer architectures. In the light of the results obtained, the Vision Transformer method has the best average loss and accuracy values compared to other methods. The accuracy and loss values of the Vision Transformer method are between 0.9871 and 0.9984 and 0.0070–0.0076, respectively. As seen in Figs. 6 and 7, the box plot of the Vision Transformer architecture is much smaller than other architectures. In addition, the distance between the extreme values in the boxplot for the Vision Transformer architecture is very small and the difference in accuracy rates is very small. It is seen that the box drawing lengths of the Vision Transformer architecture are shorter than the box drawing lengths of other architectures, the distance of the whiskers to the box is closer, and the median value is in the middle of the box. According to the results, it is seen that the Vision Transformer architecture offers more stable results in the dataset prepared for determining lemon quality compared to other architectures. To show the contribution of the Vision Transformer method, which is the most successful method we proposed in the study, to the literature, comparisons were made with the studies conducted on the same dataset and the results are shown in Table 5.

As seen in Table 5, when the studies on the quality evaluation of the lemon product were examined, the Vision Transformer model used in the study provided a higher success rate than other studies. Fruit diseases are one of the serious major problems in lemon cultivation. Therefore, the detection of these diseases is of vital importance for the cultivation of lemons and other fruits. Lemon is a fruit that is frequently consumed in many parts of the world. Since it is a potential therapeutic for diseases such as cancer and tumors, and also because the vitamins it contains are extremely important for human health, lemon quality and detection of lemon diseases is an important issue. Previously, the detection of these diseases could only be done by observation. Today, these diseases can be detected automatically with image processing methods. In this study, various deep learning methods were used to classify lemon quality. Vision Transformer and Swin Transformer methods, which are new methods in the literature, and ready-made models such as EfficientNet-B5 and DenseNet-201 were used, and the performance of these models was compared. The proposed Vision Transformer model performed better than the other models. This study makes a successful contribution to the literature for lemon quality classifications, as seen in Table 5.

Conclusion

The results obtained emphasize the importance of proper classification based on physical characteristics of agricultural products and early diagnosis of diseases. Lemon is an important agricultural crop that requires a proper classification due to its annual production cycle and its nutritional value as a rich source of vitamin C and antioxidants. Lemon production in Turkey has increased continuously over the years and has experienced the fastest growth in recent years. Of this production, 45% is used for local consumption while 54.8% is exported. Accurate grading of lemons and early detection of diseases is also very important due to their value and quality in the market. Traditional manual methods of classification and diagnosis are not only slower, more demanding and inefficient, but also run the risk that experts are easily influenced by external factors such as fatigue, experience, and psychological state. This can lead to misclassifications and diagnoses and reduce the market value of the product. To overcome these challenges, artificial intelligence and computer vision technologies are being used to increase the precision of correct classifications, reduce erroneous operations, and increase efficiency in operations. This approach also allows experts to focus more on their areas of expertise and increase the economic prosperity of the country. The results presented in this paper emphasize the importance of classification of agricultural products and highlight the achievements of artificial intelligence and deep learning methods in this field.

The lemon dataset comprises 2076 images of lemons captured on a concrete surface, which were preprocessed using image processing techniques. Data augmentation such as rescaling, random zoom, random flip, and random rotation was performed before training with transformer methods which are Vision Transformer (ViT), Swin Transformer, and deep learning methods, which are Xception, ResNet-50, InceptionV3, NASNetMobile, EfficientNet-B5, InceptionResNetV2, ResNet-152, DenseNet-201. Our transformer methods performed better than other deep learning methods. Our Vision Transformer model showed 99.84% of accuracy, and our Swin Transformer method showed success on the problem with 99.23% of accuracy. As a result of the study, transformer models have taken their place in the literature as the most successful methods of lemon quality classification. With the incorporation of the decision model proposed in this study, a versatile device can be created which is applicable in various fields such as plant cultivation, agricultural product classification, disease diagnosis, and productivity enhancement. As a result, it would provide benefits such as improved and sustainable plant growth, increased speed and accuracy in agricultural product classification, early diagnosis of plant diseases to minimize product loss, and increased product quality and yield through effective crop monitoring. In future works, we will include an agricultural engineer in the study team to perform and interpret artificial intelligence studies specific to lemon varieties. We also plan to develop real-time lemon quality decision support systems using the transformer and deep learning models proposed in this study.

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

Data availability The dataset used in the experimental stages of this work was obtained from the link https://www.kaggle.com/datasets/yusufemir/lemon-quality-dataset.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Compliance with ethics requirements This article does not contain any studies with human or animal subjects.

References

1. Şahïn G (2022) Türkiye Limon Yetiştiriciliğinin Ziraat Coğrafyası Perspektifinde Analizi. Ahi Evran Akad 3(2):54
2. P. Durgapal, D. Rana, S. Aggarwal, and A. Gautam, 'Defective Fruit Classification using Variations of GAN for Augmentation', in 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Prayagraj, India: IEEE, Dec. 2022, pp. 1–6. https://doi.org/10.1109/UPCON56432.2022.9986472.
3. Adem K (2022) Impact of activation functions and number of layers on detection of exudates using circular Hough transform and convolutional neural networks. Expert Syst Appl 203:117583. https://doi.org/10.1016/j.eswa.2022.117583
4. Hekim M, Cömert O, Adem K (2020) A hybrid model based on the convolutional neural network model and artificial bee colony or particle swarm optimization-based iterative thresholding for the detection of bruised apples'. Turk J Electr Eng Comput Sci 28(1):61–79. https://doi.org/10.3906/elk-1904-180
5. P. N, P. R. K. G, P. Chanduru N M, K. N, and N. V. Fruit Disease Classification using Convolutional Neural Network. in 2022 3rd International Conference on Electronics and Sustainable

Communication Systems (ICESC), Aug. 2022, pp. 1052–1057. https://doi.org/10.1109/ICESC54411.2022.9885440 (2022)
6. Long J, Chen Y, Yang Z, Huang Y, Li C (2022) A novel self-training semi-supervised deep learning approach for machinery fault diagnosis. Int J Prod Res. https://doi.org/10.1080/00207543.2022.2032860
7. Zhang D, Gao X (2021) Soft sensor of flotation froth grade classification based on hybrid deep neural network. Int J Prod Res 59(16):4794–4810. https://doi.org/10.1080/00207543.2021.1894366
8. Glaeser A et al (2021) Applications of deep learning for fault detection in industrial cold forging. Int J Prod Res 59(16):4826–4835. https://doi.org/10.1080/00207543.2021.1891318
9. Palombarini JA, Martínez EC (2022) End-to-end on-line rescheduling from Gantt chart images using deep reinforcement learning. Int J Prod Res 60(14):4434–4463. https://doi.org/10.1080/00207543.2021.2002963
10. Q. Cheng, J. Li, G. Shen, and Q. Du. Digital Image Soil Analysis based on Machine Learning in 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC) pp. 673–677. https://doi.org/10.1109/ICNISC54316.2021.00127 (2021)
11. Jahanbakhshi A, Kheiralipour K (2020) Evaluation of image processing technique and discriminant analysis methods in postharvest processing of carrot fruit. Food Sci Nutr 8(7):3346–3352. https://doi.org/10.1002/fsn3.1614
12. Khojastehnazhand M, Mohammadi V, Minaei S (2019) Maturity detection and volume estimation of apricot using image processing technique. Sci Hortic 251:247–251. https://doi.org/10.1016/j.scienta.2019.03.033
13. Unal Y, Taspinar YS, Cinar I, Kursun R, Koklu M (2022) Application of pre-trained deep convolutional neural networks for coffee beans species detection. Food Anal Methods 15(12):3232–3243. https://doi.org/10.1007/s12161-022-02362-8
14. Adem K, Ozguven MM, Altas Z (2023) A sugar beet leaf disease classification method based on image processing and deep learning. Multimed Tools Appl 82(8):12577–12594. https://doi.org/10.1007/s11042-022-13925-6
15. Koklu M, Ozkan IA (2020) Multiclass classification of dry beans using computer vision and machine learning techniques. Comput Electron Agric 174:105507. https://doi.org/10.1016/j.compag.2020.105507
16. Md. M. Hasan, M. U. Islam, and M. J. Sadeq. A Deep Neural Network for Multi-class Dry Beans Classification', in 2021 24th International Conference on Computer and Information Technology (ICCIT), pp. 1–5. https://doi.org/10.1109/ICCIT54785.2021.9689905(2021)
17. Avuçlu E, Taşdemir Ş, Köklü M (2023) A new hybrid model for classification of corn using morphological properties. Eur Food Res Technol 249(3):835–847. https://doi.org/10.1007/s00217-022-04181-x
18. Koklu M, Kursun R, Taspinar YS, Cinar I (2021) Classification of date fruits into genetic varieties using image analysis. Math Probl Eng 2021:e4793293. https://doi.org/10.1155/2021/4793293
19. Koklu M, Sarigil S, Ozbek O (2021) The use of machine learning methods in classification of pumpkin seeds (*Cucurbita pepo* L). Genet Resour Crop Evol 68(7):2713–2726. https://doi.org/10.1007/s10722-021-01226-0
20. Kiliçarslan S (2022) Kurum Üzüm Tanelerinin Sınıflandırılması İçin Hibrit Bir Yaklaşım. Mühendis Bilim Ve Araştırmaları Derg. https://doi.org/10.6387/bjesr.1084590
21. He Y, Zhu T, Wang M, Lu H (2021) On lemon defect recognition with visual feature extraction and transfers learning process. J Inf Data Anal. https://doi.org/10.4236/jdaip.2021.94014
22. R. Sharma and V. Kukreja. Amalgamated convolutional long term network (CLTN) model for lemon citrus canker disease multi-classification', in 2022 International Conference on decision aid sciences and applications (DASA) pp. 326–329. https://doi.org/10.1109/DASA54658.2022.9765005 (2022)
23. Hernández A, Ornelas-Rodríguez FJ, Hurtado-Ramos JB, González-Barbosa JJ (2021) Accuracy comparison between deep learning models for mexican lemon classification in telematics and computing. In: Mata-Rivera MF, Zagal-Flores R (eds) Communications in computer and information science. Springer International Publishing, Cham
24. A. Pramanik, A. Zabed Khan, A. A. Biswas, and M. Rahman. lemon leaf disease classification using CNN-based architectures with transfer learning', in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) pp. 1–6. doi: https://doi.org/10.1109/ICCCNT51525.2021.9579586 (2021)
25. Bird JJ, Barnes CM, Manso LJ, Ekárt A, Faria DR (2022) Fruit quality and defect image classification with conditional GAN data augmentation. Sci Hortic 293:110684. https://doi.org/10.1016/j.scienta.2021.110684
26. Akerlof GA (1970) The market for lemons: quality uncertainty and the market mechanism. Q J Econ 84(3):488–500. https://doi.org/10.2307/1879431
27. Lemon Quality Dataset. Accessed: 02 Nov. 2023. Available: https://www.kaggle.com/datasets/yusufemir/lemon-quality-dataset
28. M. M. Shahriar Maswood, T. Hussain, M. B. Khan, M. T. Islam, and A. G. Alharbi. CNN based detection of the severity of diabetic retinopathy from the fundus photography using efficientnet-B5 in 2020 11th IEEE Annual information technology, electronics and mobile communication conference (IEMCON), Nov pp. 0147–0150. https://doi.org/10.1109/IEMCON51383.2020.9284944 (2020)
29. S. Wu, J. Wang, Y. Ping, and X. Zhang, 'Research on Individual Recognition and Matching of Whale and Dolphin Based on EfficientNet Model', in 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), pp. 635–638. https://doi.org/10.1109/ICBAIE56435.2022.9985881 (2022)
30. R. N. Lazuardi, N. Abiwinanda, T. H. Suryawan, M. Hanif, and A. Handayani automatic diabetic retinopathy classification with EfficientNet in 2020 IEEE REGION 10 CONFERENCE (TENCON), , pp. 756–760. https://doi.org/10.1109/TENCON50793.2020.9293941 (2020)
31. M. Tan and Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv. Accessed: Nov 02, 2023. http://arxiv.org/abs/1905.11946 (2020)
32. Y. Altaf, A. Wahid, and M. M. Kirmani. Deep Learning approach for sign language recognition using DenseNet201 with Transfer Learning', in 2023 IEEE International Students Conference on Electrical, Electronics and Computer Science (SCEECS), Feb. pp. 1–6. https://doi.org/10.1109/SCEECS57921.2023.10063044 (2023)
33. P. Padhi and M. Das. Hand gesture recognition using denseNet201-mediapipe hybrid modelling in 2022 International Conference on automation, computing and renewable systems (ICACRS), Pudukkottai, India: IEEE. pp. 995–999. https://doi.org/10.1109/ICACRS55517.2022.10029038. (2022)
34. A. D. J. Abadicio et al. Ground-level Post-disaster image classification using DenseNet201 for disaster damage assessment', in 2023 International Conference On Cyber Management And Engineering (CyMaEn) pp. 132–137. doi: https://doi.org/10.1109/CyMaEn57228.2023.10050981 (2020)
35. A. Dosovitskiy et al., 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale presented at the

International Conference on Learning Representations. Accessed: Nov. 02, 2023. https://openreview.net/forum?id=YicbFdNTTy. (2020)
36. Z. Liu et al. Swin transformer: hierarchical vision transformer using shifted windows presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society. pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986 (2021)
37. Tuncel İ, Albayrak A, Akin M (2022) Öz Dikkat Mekanizması Tabanlı Görü Dönüştürücü Kullanılarak Sıtma Parazit Tespiti. DÜMF Mühendis Derg. https://doi.org/10.24012/dumf.1120289
38. M. A.-E. Zeid, K. El-Bahnasy, and S. E. Abo-Youssef. Multiclass colorectal cancer histology images classification using vision transformers, in 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt: IEEE, pp. 224–230. https://doi.org/10.1109/ICICIS52592.2021.9694125 (2021)
39. M. T. Mali, E. Hancer, R. Samet, Z. Yıldırım, and N. Nemati. Detection of colorectal cancer with vision transformers in 2022 innovations in intelligent systems and applications Conference (ASYU), pp. 1–6. https://doi.org/10.1109/ASYU56188.2022.9925335. (2022)
40. Z. Liu et al. Swin Transformer V2: Scaling up capacity and resolution', in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 11999–12009. https://doi.org/10.1109/CVPR52688.2022.01170 (2022)
41. Li L-H, Tanone R (2023) Disease identification in potato leaves using swin transformer in 2023. Int Conf Ubiquit Inform Manag Commun (IMCOM). https://doi.org/10.1109/IMCOM56909.2023.10035609
42. Powers DMW (2020) Evaluation: from precision, recall and F-measure to ROC informedness, markedness and correlation. arxiv. https://doi.org/10.4850/arXiv.2010.16061
43. N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: generalization gap and sharp minima. arXiv. Accessed 02 Nov 2023. http://arxiv.org/abs/1609.04836
44. Novakovic J, Veljovic A, Ilić S, Papic Ž, Milica T (2017) Evaluation of classification models in machine learning. Theory Appl Math Comput Sci 7(1):39
45. Ferdinandy B et al (2020) Challenges of machine learning model validation using correlated behaviour data: evaluation of cross-validation strategies and accuracy measures. PLoS ONE 15(7):e0236092. https://doi.org/10.1371/journal.pone.0236092
46. Hossin M, Sulaiman MN (2015) A Review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5(2):01–11. https://doi.org/10.5121/ijdkp.2015.5201