Vol: 3, Issue: 1, 2025
Pages:1-16
Received: 3 December 2024
Accepted: 26 December 2024

CURRENT TRENDS IN COMPUTING
Research Article
DOI: 10.71074/CTC.1595868

CLASSIFICATION OF ART PAINTINGS USING VISION TRANSFORMERS

NERGİZ İNAL[1*], SERDAR ÇİFTÇİ[1]

[1]*Department of Computer Engineering, Harran University, 63300, Şanlıurfa, Türkiye*

ABSTRACT. In this study, we investigated the performance of art painting classification based on the categories of "Genre", "Artist", and "Style" using deep learning methods. We employed convolutional neural network (CNN)-based models such as ResNet, MobileNet, EfficientNet, and ConvNeXt for their effectiveness in feature extraction. Additionally, we used vision transformer models, including ViT, Swin, BEiT, and DeiT, which used attention mechanisms. We conducted our experiments on the publicly available WikiArt dataset, and the BEiT model achieved the highest classification accuracy in the Artist and Genre categories, with results of 84.90% and 79.52%, respectively. In Style category, the Swin model produced the best result with an accuracy of 72.59%. In general, our findings indicate that transformer-based methods outperformed CNN-based methods. Furthermore, we compared our results with similar studies in the literature and showed that transformer-based models generally perform better in classifying art paintings.

1. INTRODUCTION

Classifying and categorizing art paintings accurately is important for both historical and aesthetic reasons. With the increasing digitalization of visual art today, analyzing large volumes of visual data has become more challenging. Consequently, machine learning techniques have begun to play a crucial role in this process [1]. In particular, the power of deep learning algorithms offers significant advantages in accurately classifying artworks by genre, artist, and style [2].

Although the number of studies focusing on artwork classification has risen in recent years, these studies generally concentrate on a limited number of artists or genres and often involve only a few artists for classification. Furthermore, there is a lack of recent comparative studies in the literature that highlight the performance differences between transformer-based deep learning models [3] [4] and traditional Convolutional Neural Networks (CNN) [5] across various classification levels, including artist, genre, and style.

In this study, we investigated the classification of art paintings using deep learning techniques, specifically CNN-based models and vision transformers. We compare the performance of these two architectures to determine if transformer-based models achieve higher accuracy rates than CNN-based ones in classifying art paintings. Additionally, we investigate which classification tasks—artist, style, or genre—each model performs better. By evaluating each model's performance across various classification tasks, we assess the effectiveness of these architectures in the context of art paintings. For our

E-mail address: inalnergiz@gmail.com (*), serdarciftci@harran.edu.tr.
Key words and phrases. Painting Classifications, Vision Transformers, Transfer Learning.

study, we utilized several transformer-based models, including BEiT [6], ViT [4], Swin [7], and DeiT [8]. Additionally, we employed traditional CNN architectures such as ResNet [9], MobileNet [10], EfficientNet [11], and ConvNeXt [12]. These deep learning architectures were specifically chosen for the task of classifying art paintings and were trained and tested using the WikiArt dataset [13].

FIGURE 1. **Sample predictions of the BEiT model for the "Artist" class**.

We analyzed the performance of the models using evaluation metrics such as accuracy, precision, recall, and F1 score. In Figure 1, we present the prediction results produced by BEiT, which achieved the best outcomes among the models we used, particularly for the "Artist" class in art paintings. These figures display both the ground truth labels and the predicted labels for each artwork, effectively demonstrating the classification capability of the model.

The remainder of this paper is organized as follows: Section 2 presents a literature review that examines previous studies related to the topic. In Section 3, we describe the dataset, the models utilized, and the methodologies applied. Section 4 offers our experimental results alongside a comparative analysis. Finally, Section 5 concludes the study by summarizing the key findings.

2. Related Work

The classification of paintings is a significant area of research in both art history and artificial intelligence. To achieve this, various methods have been developed to categorize paintings based on attributes such as artist, genre, and style. The literature has classified studies into three main approaches: machine learning-based, CNN-based, and transformer-based methods.

2.1. **Machine Learning-Based Studies:**

Carneiro et al. [14] introduced a comprehensive dataset comprising 988 religious-themed art prints from the 15th to 17th centuries. The study evaluated various methods for artistic image annotation and retrieval tasks, such as Bag-of-Features and Label Propagation. Notably, the innovative Inverted Label Propagation method yielded the most promising results. In another study [11], the performance of a conditional generative adversarial network (ArtGAN) model was assessed on multiple datasets, including CIFAR-10, STL-10, and Oxford-102 [15]. ArtGAN achieved state-of-the-art results on CIFAR-10, particularly excelling in terms of the Inception Score. Additionally, [16] employed various algorithms to classify painters based on artworks from the WikiArt dataset. Sandoval et al. [17] proposed an unsupervised Adversarial Clustering System (ACS) that does not require human annotation. ACS successfully labeled artworks with higher reliability through its unsupervised clustering and classification modules.

2.2. **CNN-Based Studies:**

Agarwal et al. [18] emphasized that as digitized painting collections grow larger, organizing and retrieving paintings from these collections becomes increasingly challenging. Their study focused on feature extraction and classification of paintings by genre and style. The model they developed achieved an accuracy of 84.56% in genre classification and 62.37% in style classification across six genres and ten styles. In another study [19], Deep Convolutional Networks were utilized for large-scale artwork classification. The study had two primary objectives: first, to train an end-to-end deep model for classifying artworks, and second, to assert that classifying artworks is a more complex task than object or face recognition. Using the large WikiArt dataset, their proposed model achieved a success rate of 68%. Jangtjik et al. [20] proposed a CNN-LSTM-based model to address the challenges of artist classification in digital paintings, performing multi-class classification among 13 artists by analyzing local patch information. The study [21] aimed to detect artistic styles using the ResNet-50 architecture and sought to outperform existing methods, achieving 62% accuracy on the WikiArt dataset across 25 different styles. Cetinic et al. [22] examined the applicability of CNNs for classification tasks in the realm of art paintings, attaining 81.94% accuracy in artist classification, 77.60% in genre classification, and 56.43% in style classification. Chu et al. [23] explored deep correlation features for the style classification of artworks, using a dataset of 19,787 oil paintings spanning 17 styles, all collected from WikiArt. In [24], the effects of deep learning and transfer learning methods on image classification were examined. The research focused on several popular CNN architectures, including LeNet, AlexNet, GoogleNet, VGG16, VGG19, and ResNet-50, along with the application of transfer learning techniques.

2.3. **Transformer-Based Studies:**

Conde et al. [25] highlighted the challenges of recognizing fine details in artworks and the significant costs associated with creating labeled datasets. In their study, a neural network was trained on pairs

of art images and their corresponding text descriptions using CLIP (Contrastive Language-Image Pre-Training). The model successfully predicted the most suitable natural language description for an image through its zero-shot learning capability, eliminating the need for prior training. The study [26] compared the efficiency of CNN-based methods with transformer-based methods, specifically Vision Transformers (ViT). Experiments carried out on a butterfly dataset consisting of 10,000 data points revealed that the ViT model delivered more accurate results with larger datasets, although it required a more costly and time-consuming training process. The study emphasized the potential for improvement in the ViT model's application in industrial settings. The study by Iliadis et al. [27] employed two deep learning architectures, ViT and MLP-Mixer, for artwork recognition. Both models were trained from scratch and achieved an accuracy of over 39% on 21 style classes using the WikiArt dataset. In [28], the performance of ViT and ResNet-50 was compared on a small art history dataset. Using a large portrait collection spanning from the 15th to the 19th century, the study demonstrated that ViT outperformed ResNet-50, achieving an accuracy of 87.09% compared to 46.13%. Schaerf et al. [29] investigated the use of vision transformers in the task of artwork authentication. In a dataset containing both original paintings by Vincent van Gogh and imitations, the study revealed that the Swin Transformer model outperformed EfficientNet, achieving an accuracy of over 85%.

3. Materials and Methods

3.1. **Dataset.**

We used the WikiArt dataset [13], the largest accessible online archive of digitized artworks, for our experiments. The collection has been meticulously curated to include various metadata attributes such as artist information, artistic style, and genre categorization.

3.2. **Distribution of the WikiArt Dataset.**

The WikiArt dataset contains 80,735 different artworks. For our experiments, we used WikiArt subsets that included artists, styles, and genre classes. A subset of 23 artists with at least 850 paintings was selected for artist classification. A subset of 15 different classes was selected for style classification, each containing more than 900 images. For genre classification, a subset consisting of 10 classes was selected, each with more than 1,880 paintings. Table 1 presents the class distributions. The distributions for each class of artist, genre, and style are shown in Figure 2, and some sample images from the WikiArt dataset are presented in Figure 3.

Table 1. **Class-wise distribution of the WikiArt subsets**

Category	Number of Classes	Number of Images
Artist	23	43,424
Genre	10	18,800
Style	15	13,500

3.3. **Data Augmentation.**

During the data preprocessing stage, data augmentation techniques were applied to address class imbalance and to enable the model to learn in a more generalized and robust manner. As part of the augmentation process, various random transformations were applied to the images to diversify the dataset

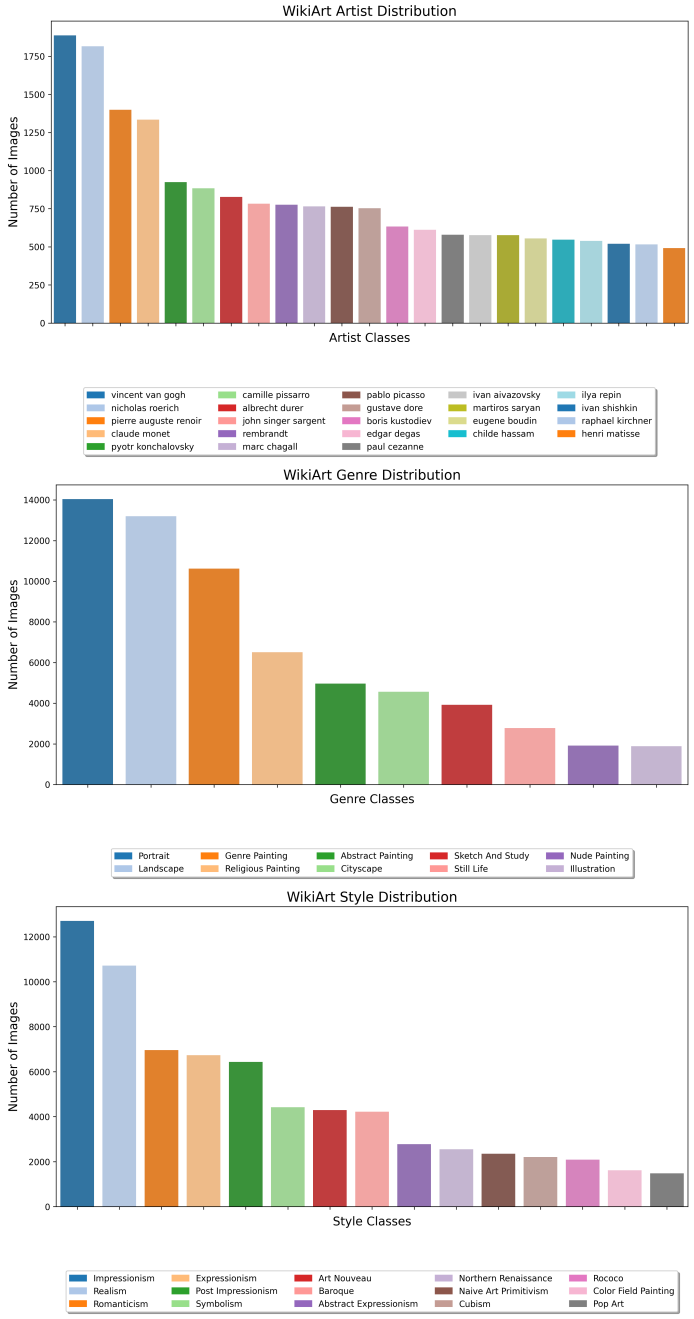

FIGURE 2. **The distribution of painting classes with respect artist, genre, and style.**

and enrich it for learning under different conditions. The applied augmentation techniques included random 90-degree rotation (RandomRotate90), horizontal flipping (HorizontalFlip), brightness and contrast

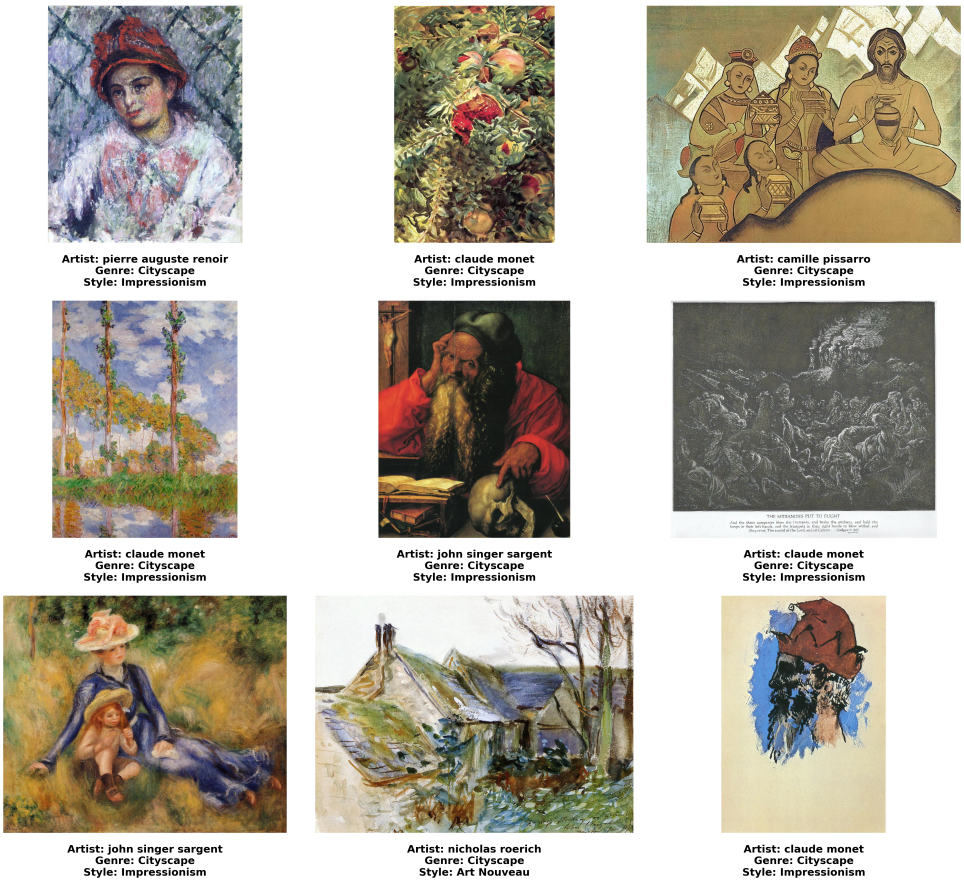

FIGURE 3. **Sample images from the WikiArt dataset.**

adjustment (RandomBrightnessContrast), Gaussian blur (GaussianBlur), and hue-saturation-value adjustment (HueSaturationValue). These operations were specifically performed to reduce class imbalance and to allow the model to be trained on a more varied dataset.

3.4. **Methods:**

This section provides detailed explanations of the models used. The models employed for painting classification are examined under two main categories: transformer-based and CNN-based models. Transformer-based models offer the capacity to learn complex relationships through attention mechanisms [3], while CNN-based models can effectively extract visual features through convolutional layers [5].

3.4.1. *Transformer-Based Models:*

1.Vision Transformer (ViT).

The ViT [4] is a deep learning model that applies the transformer architecture, originally developed for natural language processing tasks, to visual data. Instead of processing the image at the pixel level,

ViT divides the image into fixed-size small patches and flattens each patch into a vector. These vectors are then used as input to the transformer model, where the relationships between the patches are learned through attention mechanisms. In this way, ViT interprets the overall structure and content of the image by considering the contextual relationships between the patches.

2. Swin Transformer.

The Swin Transformer [7] adopts a more efficient approach for processing visual data. This efficiency is achieved by dividing images into small local patches and processing each patch through dedicated attention layers. Initially, large images are split into small patches of 16×16 pixels. Each patch learns internal relationships through local attention layers. By using the sliding window technique, each patch attempts to understand not only its own internal structure but also its relationships with neighboring patches. Due to the hierarchical design, as the model progresses from small-scale to larger-scale patches, global attention layers are activated to capture relationships over broader regions.

3. DeiT (Data-efficient Image Transformer).

DeiT [8] is a variant of the ViT model, specifically designed to achieve high performance with smaller datasets. Similar to ViT, this model processes images by dividing them into small patches. Each patch is converted into a vector that carries its features and is fed into the model. The main distinction from ViT lies in DeiT's use of the Teacher-Student learning approach. In this framework, a large pre-trained model (the teacher) guides the training of a smaller model (the student). The teacher model transfers correct classification knowledge to the student model, which learns by mimicking the teacher's outputs.

4. BEiT (BERT Pretrained Image Transformer):

BEiT [6] is a model that leverages natural language processing (NLP) techniques to process visual data. The image is typically divided into small patches of 16×16 pixels, and each patch is converted into a vector that represents the features the model can understand. These vectors are fed into the model, and relationships between the patches are learned through attention mechanisms. BEiT adapts the masked language modeling approach from the BERT model to visual data by masking certain image patches and attempting to predict the missing parts. This allows the model to learn the global context of the image more effectively.

3.4.2. *CNN-Based Models:*

1. ResNet (Residual Networks).

ResNet [9] is a CNN model designed to overcome the challenges of training very deep networks in deep learning. ResNet utilizes a "residual learning" structure, where each layer directly uses the output from the previous layer. These direct connections are known as shortcut connections. Shortcut connections accelerate the learning process between layers, allowing effective learning even in deeper layers of the network. The main advantage of ResNet is that it eliminates the vanishing gradient problem encountered during the learning process, even in very deep networks.

2. MobileNet.

MobileNet [10] is a CNN-based deep learning model optimized for use on hardware with limited computational power, such as mobile devices and embedded systems. MobileNet's most important feature is its use of the depthwise separable convolution approach, which minimizes traditional convolutional

layers. This approach performs convolution operations separately for each channel, requiring less computational power compared to traditional convolutional layers, allowing the model to operate faster and be more suitable for mobile devices' processing power and memory capacity.

3. EfficientNet.

EfficientNet [11]is an efficient CNN architecture developed by Google, which provides high accuracy in tasks such as image classification. It delivers high performance with fewer parameters by using the Compound Scaling method, which scales depth, width, and input size in a balanced way. Based on the MobileNetV2 structure, EfficientNet is both a powerful and lightweight model family suitable for mobile devices.

4. ConvNeXt.

ConvNeXt [12] is an image classification model that updates the traditional Convolutional Neural Networks (CNN) architecture with modern deep learning techniques. Inspired by the ViT model, ConvNeXt modernizes the CNN structure with components like large kernels, GELU activation, and Layer Normalization. This approach maintains computational efficiency while achieving strong performance with high accuracy.

3.5. **Hyperparameter Selection:**

Training parameters are one of the most important factors that directly affect the performance of the model. To determine the optimized parameters, we used the Optuna[1] framework, which uses Bayesian optimization to efficiently search for the model's hyperparameters. This process selects each parameter trial based on previous results, allowing the model to reach the optimal parameter combinations with fewer trials. The TPE (Tree-structured Parzen Estimator) algorithm predicts which parameter combinations will yield better results based on data from past trials, enabling more targeted selections in each new trial. The optimization process was carried out with a trial number (`n_trial`) set to 3, and the most suitable hyperparameters were determined. During the optimization, parameters such as batch size (the number of samples the model processes at once during training), learning rate (which defines the step size for weight updates), epochs (the number of training cycles over the entire dataset), and warmup steps (which control how quickly the learning rate increases at the start of training) were considered. Additionally, the AdamW optimizer was used to facilitate more efficient learning and achieve better results. The parameters used in this study are listed in Table 2 to ensure efficient model training.

TABLE 2. **Ranges of hyperparameters used for model training**

Hyperparameter	**Value Range**
Batch Size	16, 32, 64
Learning Rate	1e-5, 5e-4
Epochs	20 - 30
Warmup Steps	100 - 200

[1] `www.optuna.org`

4. EXPERIMENTS

4.1. **Model Training Process:**

4.1.1. *Experimental Environment.*

We used the TRUBA[2] computing infrastructure for model training. During the training process, the system was equipped with 50 GB of system RAM, one CUDA-supported NVIDIA GPU, and a 20-core CPU. The software configuration included the RockyLinux-9.2 operating system, Python 3.x, and the PyTorch deep learning library for model training, along with Hugging Face Hub for managing and deploying the models.

4.1.2. *Performance Evaluation Metrics.*

A comprehensive assessment was carried out to assess the performance of the model, considering various accuracy and error metrics. The metrics used in this evaluation allowed for analyzing the model's classification success, generalization capability, and discriminative power across different classes. The results obtained were used to compare the model's performance in both the training and test datasets. The used metrics are explained as follows.

- **Accuracy:** It is the ratio of the correctly predicted examples to the total number of predictions is called accuracy. It is one of the fundamental metrics used to evaluate the overall performance of trained models. The accuracy is calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

  where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative predictions, respectively.

- **Precision:** It is a metric that measures the accuracy of the model's positive predictions. It indicates how many of the predicted positive samples are actually positive. The precision is computed as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

- **Recall:** It is a metric that measures how well the model identifies positive samples. It shows how well the model classifies the true positive examples. The recall is calculated as follows:

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

- **F1-Score:** It is a metric that balances Precision and Recall. It is particularly useful when one of these values (Precision or Recall) is low. The F1-Score is computed as follows:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

[2] `https://www.truba.gov.tr/`

4.2. **Experimental Results:**

The performance of various deep learning models (Beit, Swin, ViT, DeiT, ResNet, MobileNet, EfficientNet, and ConvNext) that evaluated on artist, genre, and style classification tasks are shown Table 3. and Table 4. The performance of the transformer-based models BEiT and Swin was compared with ViT and DeiT on the WikiArt dataset [13]. ViT and DeiT performed lower compared to the other models. Overall, BEiT and Swin provided higher accuracy and recall rates. BEiT achieved the highest performance in the "Artist" classification task with an accuracy of 84.90%, while Swin showed the best performance in the "Style" classification task with an accuracy of 72.59%.

TABLE 3. **The performance results of the transformer-based models obtained in this study**

Model	Class	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
BEiT	Artist	**84.90**	**84.79**	**84.83**	**84.90**
	Genre	**79.52**	**79.35**	**79.31**	**79.52**
	Style	68.07	68.27	68.91	68.07
Swin	Artist	82.80	82.75	82.88	82.80
	Genre	76.44	76.20	76.12	76.44
	Style	**72.59**	**72.73**	**73.23**	**72.59**
ViT	Artist	82.02	81.98	82.02	82.02
	Genre	75.96	75.86	75.93	75.96
	Style	68.07	68.21	68.47	68.07
DeiT	Artist	79.78	79.69	79.69	79.78
	Genre	73.51	73.47	73.57	73.51
	Style	68.52	68.42	68.82	68.52

Table 5 presents the results of hyperparameter optimization performed using the Optuna library to improve the model's performance. During this process, various parameter combinations were evaluated for the models, and the best results were selected based on their success on the validation dataset. These parameters were used during the training phase to prevent overfitting and to enhance the overall performance of the models.

The BEiT model achieved the best performance in genre classification, and we presented the confusion matrix in Figure 4. The value of 175 in the second row and second column reveals that the model most accurately classified artworks in the Cityscape genre. This indicates the model's strong performance in correctly identifying artistic genres. Lastly, the value of 27 in the fourth row and eighth column represents the misclassifications in the Illustration genre. The most frequent misclassification occurred in the Illustration genre, while the highest number of correct classifications was observed in the Cityscape genre.

Table 6 presents a comparative overview of the classification performances of several recent methods from the literature. The multi-scale CNN-based approach proposed by [30] achieved an accuracy of

TABLE 4. **The performance results of the CNN-based models obtained in this study**

Model	Class	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
ResNet	Artist	76.26	76.09	76.49	76.26
	Genre	68.99	68.45	68.49	68.99
	Style	58.07	56.81	58.12	58.07
MobileNet	Artist	68.40	68.01	68.53	68.40
	Genre	69.52	69.50	69.65	69.52
	Style	55.33	55.43	58.51	55.33
EfficientNet	Artist	77.20	77.09	77.67	77.20
	Genre	71.06	70.64	70.76	71.06
	Style	63.93	63.41	63.92	63.93
ConvNeXt	Artist	**79.94**	**79.84**	**79.95**	**79.94**
	Genre	**73.72**	**73.50**	**73.50**	**73.72**
	Style	**69.26**	**69.24**	**69.66**	**69.26**

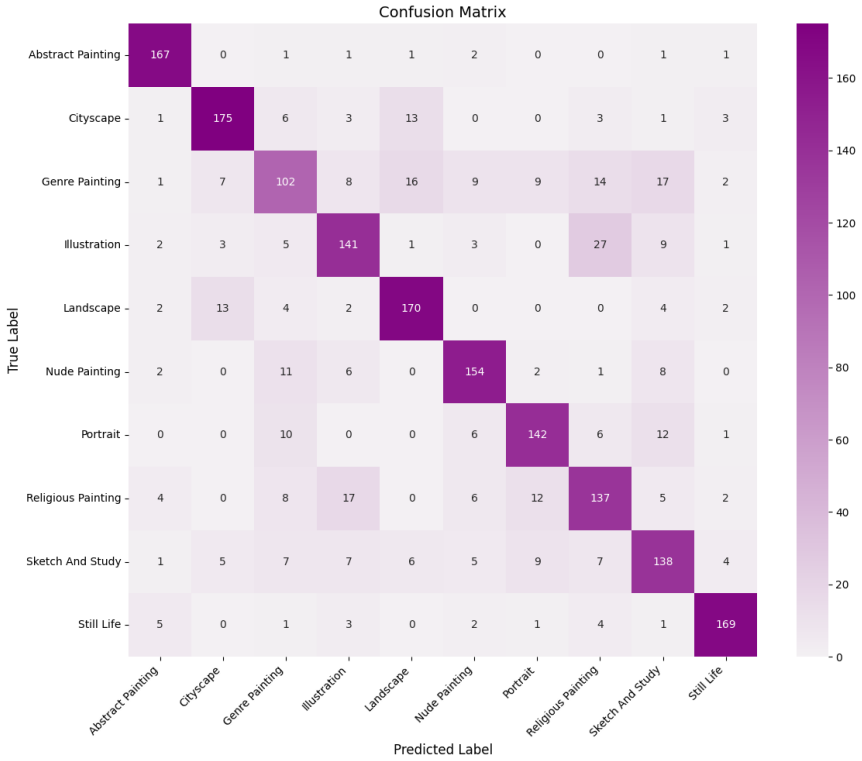

FIGURE 4. **Confusion matrix of the BEiT model for genre classification.**

79.11%. Although this method offers certain advantages over traditional CNN architectures, it demonstrated lower performance compared to the transformer-based methods used in our study. The BEiT

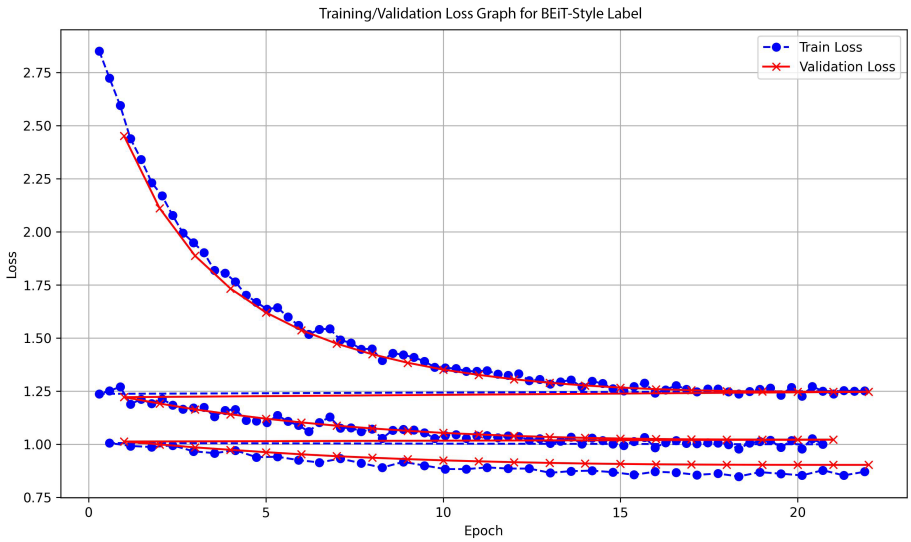

FIGURE 5. **Changes in training and validation losses over epochs.**

TABLE 5. **Best hyperparameters for transformer-based and CNN-based models**

Model	Learning Rate	Batch Size	Epoch	Warmup Steps
BEiT	1.4983	16	21	189
ViT	0.00018	32	28	129
Swin	0.00026	16	26	101
DeiT	2.3096	32	23	123
ResNet-50	0.00011	16	30	103
MobileNet	0.00026	64	22	122
ConvNeXt	0.00021	32	24	105
EfficientNet	0.00045	32	20	105

model achieved the highest accuracy of 84.90%, outperforming other approaches. The Swin transformer ranked second with an accuracy of 82.80%, following BEiT. These results indicate that transformer-based approaches provide higher accuracy in art painting classification tasks compared to traditional methods.

Figure 5 illustrates the changes in training and validation losses of the BEiT model over the epochs. It can be seen that the training loss consistently decreased, while the validation loss started to increase after a certain point.

Table 7 presents comparative results of accuracy rates for different deep learning-based methods used in the classification of artwork style, artist, and genre. The table generally demonstrates a significant improvement in the classification performance of art images when transitioning from traditional CNN architectures to transformer-based models. The CNN-LMNN method proposed by [2] achieved accuracy

TABLE 6. **Comparison of BEiT and Swin models with recent studies on the artist class of the WikiArt dataset**

Method	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Multi-scale CNN [30]	77.65	65.35	68.73	79.11
BEiT	**84.83**	**84.90**	**84.79**	**84.90**
Swin	82.88	82.80	82.75	82.80

TABLE 7. **Comparison of BEiT and Swin models with recent studies based on style, artist, and genre classes**

Method	Class	Accuracy (%)
CNN-LMNN [2]	Style	45.97
	Artist	63.06
	Genre	58.48
Fine-tuning AlexNet [19]	Style	54.50
	Artist	76.11
	Genre	74.14
CNN fine-tuning CaffeNet [22]	Style	56.43
	Artist	81.94
	Genre	77.60
BEiT	Style	68.07
	Artist	**84.90**
	Genre	**79.52**
Swin	Style	**72.59**
	Artist	82.80
	Genre	76.44

rates of 45.97% for style classification, 63.06% for artist classification, and 58.48% for genre classification. Following this, the fine-tuning AlexNet model [19] reached higher accuracy rates of 54.50% for style, 76.11% for artist, and 74.14% for genre classifications. The CNN fine-tuning CaffeNet approach by [22] notably excelled in artist 81.94% and genre 77.60% classifications, while achieving an accuracy rate of 56.43% in style classification.

In our study, the results of transformer-based models were examined, revealing that these architectures achieve higher accuracy rates compared to traditional methods. The BEiT model obtained the most successful results with an accuracy rate of 84.90% in artist classification, while reaching 79.52% accuracy in genre classification and 68.07% in style classification. The second-best transformer model, Swin, demonstrated the highest performance in style classification with an accuracy rate of 72.59%. These results indicate that transformer-based models can learn more effective representations in complex and context-rich datasets such as art images, thereby achieving higher accuracy in classification tasks.

5. Conclusion

In this study, we observed that using vision transformer models to identify artists, styles, and genre classifications of paintings outperformed traditional CNN-based methods. According to the results presented in Table 6 and Table 7, the BEiT model achieved the highest accuracy in artist classification with a score of 84.90%. This indicates that the model was able to capture the characteristic drawing features and stylistic distinctions of artists with high precision. The consistent performance of the BEiT model across balancing metrics such as precision, recall, and F1-Score demonstrates that the model was effective not only in certain classes but also across the overall distribution. The Swin transformer model, with an accuracy of 82.80%, delivered performance close to that of BEiT and yielded strong results especially in recall and F1-score metrics. Furthermore, as shown in Table 6, the relatively lower performance of the Multi-scale CNN method [30] compared to transformer-based models indicates that traditional CNN architectures fall short in capturing complex stylistic elements and contextual relationships. When examining the literature comparisons presented in Tables 6 and 7, it can be observed that earlier studies—such as the CNN-LMNN method by [2]—achieved limited accuracy rates, with 45.97% in style classification. More recent CNN-based approaches [19] [22] showed partial improvements, but remained below the accuracy levels reached by the transformer-based models used in our study. In particular, the Swin model that achieves an accuracy of 72.59% even in the abstract category of style classification highlights the significant advantage of transformer architectures in understanding the complex structural characteristics of art. In this context, it is concluded that transformer-based models have greater potential in tasks like art classification, which require the simultaneous processing of visual and contextual information.

Declarations

- **Contribution Rate Statement:** All authors contributed equally to the design, implementation, analysis, and writing of this study.
- **Conflict of Interest:** The authors declare that there is no conflict of interest regarding the publication of this paper.
- **Data Availability:** The WikiArt dataset used in this study is publicly available and can be accessed through the link provided in reference [13].
- **Statement of Support and Acknowledgment:** The authors thank TRUBA for providing the necessary computational resources for this study.

References

[1] M. A. Özdal, Resim içeriği sınıflandırmasında yapay zekanın rolü, D-Sanat 1 (9) (2025) 56–71.

[2] B. Saleh, A. Elgammal, Large-scale classification of fine-art paintings: Learning the right metric on the right feature, arXiv preprint arXiv:1505.00855 (2015).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[5] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[6] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254 (2021).
[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International conference on machine learning, PMLR, 2021, pp. 10347–10357.
[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).
[11] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.
[13] W. R. Tan, C. S. Chan, H. Aguirre, K. Tanaka, Improved artgan for conditional synthesis of natural image and artwork, IEEE Transactions on Image Processing 28 (1) (2019) 394–409. doi:10.1109/TIP.2018.2866698.
URL https://doi.org/10.1109/TIP.2018.2866698
[14] G. Carneiro, N. P. Da Silva, A. Del Bue, J. P. Costeira, Artistic image classification: An analysis on the printart database, in: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12, Springer, 2012, pp. 143–157.
[15] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images.(2009) (2009).
[16] M. D. Choudhury, Automated identification of painters over wikiart image data using machine learning algorithms, Ph.D. thesis, Dublin, National College of Ireland (2020).
[17] C. Sandoval, E. Pirogova, M. Lech, Adversarial learning approach to unsupervised labeling of fine art paintings, IEEE Access 9 (2021) 81969–81985.
[18] S. Agarwal, H. Karnick, N. Pant, U. Patel, Genre and style based painting classification, in: 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2015, pp. 588–594.
[19] W. R. Tan, C. S. Chan, H. E. Aguirre, K. Tanaka, Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification, in: 2016 IEEE international conference on image processing (ICIP), IEEE, 2016, pp. 3703–3707.
[20] K. A. Jangtjik, T.-T. Ho, M.-C. Yeh, K.-L. Hua, A cnn-lstm framework for authorship classification of paintings, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 2866–2870.
[21] A. Lecoutre, B. Negrevergne, F. Yger, Recognizing art style automatically in painting with deep learning, in: Asian conference on machine learning, PMLR, 2017, pp. 327–342.
[22] E. Cetinic, T. Lipic, S. Grgic, Fine-tuning convolutional neural networks for fine art classification, Expert Systems with Applications 114 (2018) 107–118.
[23] W.-T. Chu, Y.-L. Wu, Image style classification based on learnt deep correlation features, IEEE Transactions on Multimedia 20 (9) (2018) 2491–2502.
[24] S. T. Krishna, H. K. Kalluri, Deep learning and transfer learning approaches for image classification, International Journal of Recent Technology and Engineering (IJRTE) 7 (5S4) (2019) 427–432.
[25] M. V. Conde, K. Turgutlu, Clip-art: Contrastive pre-training for fine-grained art classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3956–3960.

[26] K. Lu, Y. Xu, Y. Yang, Comparison of the potential between transformer and cnn in image classification, in: ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application, VDE, 2021, pp. 1–6.
[27] L. A. Iliadis, S. Nikolaidis, P. Sarigiannidis, S. Wan, S. K. Goudos, Artwork style recognition using vision transformers and mlp mixer, Technologies 10 (1) (2021) 2.
[28] S. Diem, T. Mandl, Automatic classification of portraits: Application of transformer and cnn based models for an art historic dataset., in: LWDA, 2023, pp. 192–206.
[29] L. Schaerf, E. Postma, C. Popovici, Art authentication with vision transformers, Neural Computing and Applications 36 (20) (2024) 11849–11858.
[30] Q. Yu, C. Shi, An image classification approach for painting using improved convolutional neural algorithm, Soft Computing 28 (1) (2024) 847–873.