Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on AI-Driven Creativity and Intelligent Practices in Visual Arts 2025 6(4s)

EXPLAINABLE AI FOR STYLE INTERPRETATION IN CONTEMPORARY ART USING VISION TRANSFORMERS

Nikil Tiwari [1], Naman Soni [2], Rahul Anantrao Padgilwar [3], Dr Preeti Pandurang Kale [4], Dr. Mandeep Kaur [5], Dr. Vinay Nagalkar [6]

[1] Assistant Professor, School of Fine Arts, AAFT University of Media and Arts, Raipur, Chhattisgarh-492001, India
[2] Assistant Professor, School of Fine Arts and Design, Noida International University, Noida, Uttar Pradesh, India
[3] Assistant Professor, Department of Desh, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India
[4] Assistant Professor, Department of Electronics and Computer Engineering, CSMSS Chh. Shahu College Of Engineering, Chhatrapati Sambhajinagar, Maharashtra, India
[5] Department of Computer Science and Engineering, CT University Ludhiana, Punjab, India
[6] Department of E and TC, Ajeenkya DY Patil School of Engineering, Pune, Maharashtra, India

ABSTRACT

This study looks into how Explainable AI (XAI) can be used to figure out style in modern art by using Vision Transformers (ViTs). As the need for interpretability in AI-driven art research has grown, models have had to be made that not only work well but also give clear, understandable reasons for the choices they make. We use ViTs, a cutting-edge deep learning system that is known for being very good at classifying images, to look at and figure out what the style aspects of modern art mean. The study aims to find a balance between the need for high-performance AI models and the need for openness in the art world. It will do this by showing how certain aspects of artworks, like colour schemes, structural structures, and brushstroke patterns, affect the overall style. We present a mixed framework that blends the power of Vision Transformers with techniques for explaining things like Grad-CAM and focus maps. This framework helps you see and understand how the model's predictions work. The results show that the model can correctly spot important creative traits and give visual descriptions, which helps people understand different types of art. Additionally, the suggested method is tried on a wide range of modern artworks, showing that it can be used with various types of art. This work has effects beyond just analysing art; it gives managers, artists, and students a useful tool for working with AI systems in a more open way. It also adds to the field of explainable AI by using these methods to study art analysis, which is very biassed and hard to explain.

Keywords: Explainable AI, Vision Transformers, Style Interpretation, Contemporary Art, Attention Maps

Received 12 May 2025
Accepted 14 September 2025
Published 25 December 2025

Corresponding Author
Nikil Tiwari, nikhil.tiwari@aaft.edu.in

DOI
10.29121/shodhkosh.v6.i4s.2025.6941

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

Numerous interests have been paid to the area in which synthetic Genius (AI) and art meet in recent years. Deep learning models and different AI technology have proven remarkable capabilities in each analysing and making art. One huge hassle with AI-driven artwork studies, even though, is that the fashions' selection-making techniques are not continually clear or clean to understand. At the same time as contemporary AI structures offer latest pace, they are regularly challenging to understand due to the fact they're so complex. This problem is especially difficult when it comes to artwork, wherein private perspectives and the creative system are very important to what the work is worth and what it skill. Owing to this, Explainable AI (XAI) has come to be an essential need, particularly for makes use of where trust and appreciation are very important Ahmed et al. (2023). The main focuses of this observe is on how XAI may be used to understand fashion in current art. There are plenty of one-of-a-kind styles, tools, and methods used in modern-day artwork, which makes it hard for AI models to understand and analyse. Cutting-edge art could be very special and abstract, which makes it an interesting challenge to examine how AI can help now not only perceive art styles but additionally give clear motives for a way they're interpreted Thampi (2022). To apprehend a style, which involves finding precise aspects of art like brushstrokes, coloration palettes, and format techniques, you need a version that could choose up on small visible clues and flip them into beneficial information.

Vision Transformers (ViTs), a brand new form of deep studying layout based totally on self-attention approaches, have currently beaten conventional convolutional neural networks (CNNs) in a number of requirements for photo popularity obligations. As an end result, ViTs are ideal for tasks that want the model to understand long-time period connections and interactions in photos. This makes them best for looking on the fantastic functions of current artwork Marcinkevičs and Vogt (2023). CNNs work by the usage of nearby receptive fields and hierarchical function extraction. ViTs, then again, have a look at the entire photo as a sequence of patches, which lets them higher apprehend global context and excessive-level features. This skill is very essential for perception current artwork's complex and frequently vague visual language. The interpretability of ViTs remains a huge trouble, despite the fact that they do thoroughly. This is in which AI strategies that could explain matters are available accessible Burkart and Huber (2021), Alayrac et al. (2022). On this study, we need to discover to connect the brilliant things that ViTs can do with the want for openness in AI artwork studies Gemini Team, Google (2023).The main goal is to come up with an comprehensible machine that makes use of ViTs and XAI strategies to present beneficial fashion readings of modern-day art. We will test the model's ability to spot important style elements like brushstroke patterns, colour use, and the structure of the composition, and then show how these views are based on the images. We hope that this method will not only give you a good AI-powered tool for analysing art, but it will also help you learn more about how AI understands different types of art. Additionally, this study has bigger effects on using AI that can explain things in areas that need human understanding and judgement, like the arts and academic art analysis.

2. LITERATURE REVIEW

2.1. OVERVIEW OF EXISTING APPROACHES TO ART STYLE INTERPRETATION

Interpreting art styles has always been a subjective task that requires knowledge of art history, visual culture, and aesthetics in order to understand the subtleties of works of art. In the past, art experts used both visual study and knowledge of the time periods during artists' lives and movements to figure out what styles meant. But since machine learning and computer vision came along, automatic ways of recognising and interpreting art styles have become more popular Colliot (2023), Zini and Awad (2022). Figure 1 shows an overview of existing approaches to art style interpretation. Early efforts to automatically recognise art styles relied on simple visual cues like colour histograms, patterns, and edge identification to put works of art into broad groups.

Figure 1

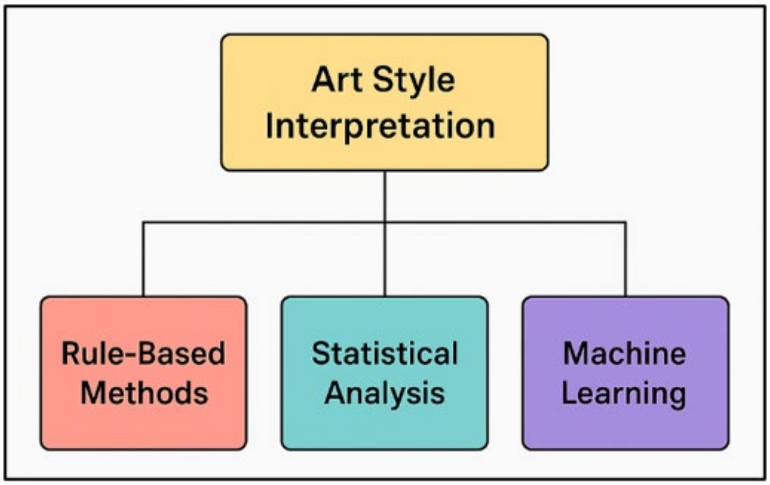

Figure 1 Overview of Existing Approaches to Art Style Interpretation

These methods worked well for some simple jobs, but they were too shallow to help me understand more complicated and abstract art. Recently, deep learning models, especially convolutional neural networks (CNNs), have been used to find more complicated things in artworks, like arrangement, brushstrokes, and patterns Vijayakumar (2022).

2.2. REVIEW OF AI AND MACHINE LEARNING APPLICATIONS IN ART

In the past few years, the use of AI and machine learning in art has grown quickly. This is due to improvements in deep learning techniques and easier access to big datasets. Machine learning models have been used in many areas of art, from making art and transferring styles to analysing and collecting art Madan et al. (2024). Deep learning, in particular, has been very helpful in handling jobs that used to need human knowledge because it can handle complex visual data. Generative Adversarial Networks (GANs) are one of the best known ways that AI is used in art to make new art. A generator and a discriminator are the two neural networks that make up a GAN. They work together to make realistic pictures that look like current art styles. Table 1 summarizes literature review with study, dataset, AI technique, findings. People have used these networks to make completely new works of art that look exactly like works made by humans.

Table 1

Table 1 Summary of Literature Review

Study/Method	Dataset Used	AI Technique	Key Findings
Gatys et al. (Neural Style Transfer)	WikiArt, ImageNet	Convolutional Neural Networks (CNN)	Pioneering style transfer, applied to paintings and photos
Johnson et al. (Perceptual Losses) Jakobsen et al. (2023)	ImageNet	Convolutional Neural Networks (CNN)	Introduced perceptual loss functions for style transfer
Li et al. (Art RecognitionYang et al. (2023)	Google Arts and Culture	ResNet, CNN	Employed Grad-CAM for explaining CNN-based style classification
Zhang et al. (Style Transfer with ViT) Yu and Xiang (2023)	WikiArt	Vision Transformer (ViT)	Introduced ViT for art style transfer with explainability via Grad-CAM
Tancik et al. (Learning Visual Styles Chan et al. (2023)	WikiArt, Art1000	CNN	Implemented SHAP to explain CNN decisions in art classification

3. METHODOLOGY

3.1. DATA COLLECTION: DESCRIPTION OF CONTEMPORARY ART DATASETS

Choosing the right modern art dataset is very important for teaching a Vision Transformer (ViT) model how to understand art styles. The WikiArt dataset is one of these. It has more than 100,000 works of art from a wide range of time periods, styles, and types, making it a useful tool for teaching deep learning models. The ArtBench dataset is another popular pick. It is a collection of modern works of art in a variety of styles and media, such as paintings, sculptures, and digital art. Usually, these sets of data include both the picture data and the information that goes with it, like the names of the artists, the art movement, and the year the work was made. In the case of style interpretation, the information needs to be clearly labelled with the styles or characteristics that the model needs to learn.

3.2. PREPROCESSING OF ART IMAGES FOR VIT TRAINING

A very important step in getting art pictures ready for Vision Transformer (ViT) training is to process them first. In their original state, raw pictures might have noise, traits that aren't important, or errors that could make the ViT model work less well. When preprocessing is done right, the model can focus on the most important visual details that are needed to understand style. As a first step in preparation, the pictures are resized so that they all come in at the same size.

1) Resizing the Image:

Given an input image I ∈ ℝ^(H × W × C), where H is the height, W is the width, and C is the number of color channels (e.g., RGB, C=3), the image is resized to a target size S (e.g., 224 × 224). This is typically done using bilinear interpolation:

$$I_{resized} = Resize(I, S)$$

where S is the desired image dimension.

2) Normalization:

The pixel values of the image I_resized are normalized to have zero mean and unit variance using the dataset's global mean and standard deviation, μ and σ, respectively:

$$I_{normalized} = \frac{(I_{resized} - \mu)}{\sigma}$$

where μ ∈ ℝ^C and σ ∈ ℝ^C are the mean and standard deviation of each channel (Red, Green, Blue).

3) Patch Extraction:

The image I_normalized is divided into non-overlapping patches of size P × P, where P is the patch size. The total number of patches is N = H / P × W / P. Each patch is flattened into a vector of size P^2 × C:

$$I\_patches = \{Flatten(I\_normalized[i, j]) \mid 1 \leq i \leq N, 1 \leq j \leq N\}$$

4) Patch Embedding:

The flattened patches are projected into a higher-dimensional space via a linear projection W_patch∈ ℝ^((P^2 × C) × D), where D is the embedding dimension:

$$z_i = I_{patch[i]} \cdot W_{patch} + b$$

wherez_i is the embedded vector for patch i, and b is the bias term.

3.3. VISION TRANSFORMER MODEL ARCHITECTURE AND TRAINING PROCESS

The Vision Transformer (ViT) design is a deep learning model that has done very well at tasks like art style analysis and picture classification. ViTs work differently than regular Convolutional Neural Networks (CNNs). They work by breaking a picture into fixed-size pieces that are then put together in a straight line to make a flat sequence. This method lets the model see how things in an image are connected and how they fit into the bigger picture. This makes it great for jobs that need a lot of visual analysis, like figuring out art styles. The ViT design is made up of several important parts. First, the picture is broken up into parts that don't touch each other. These patches are usually 16x16 or 32x32 pixels in size. The patches are then put into vectors in a straight line, and a stack of transformer layers work on them.

1) Input Embedding

Given an input image with N patches, the patches are embedded into a D-dimensional space using a linear projection. The resulting embedding for each patch z_i∈ ℝ^D is:

$$Z = [z_1, z_2, \dots, z_N] \in \mathbb{R}^{N \times D}$$

A special token z_0 is added to the sequence to represent the image's global representation.

2) Self-Attention Mechanism

The ViT model applies multi-head self-attention to the patch embeddings. For each attention head h, the attention weight is computed as:

$$Attention_h = softmax\left(\frac{(Q\,W_Q \cdot K\,W_K^T)}{\sqrt{D}}\right) \cdot V\,W_V$$

where Q, K, V are the query, key, and value matrices, and W_Q, W_K, W_V are learnable weights for each head.

3) Transformer Layer

The output of the self-attention mechanism is passed through a feed-forward neural network (FFN) layer.

Output = FFN(Attention_h) + Residual Connection

The FFN consists of two linear layers with a ReLU activation in between:

$$FFN(x) = ReLU(x\,W_1 + b_1)W_2 + b_2$$

4) Output Layer and Training Objective

After passing through several transformer blocks, the final output vector for the special token z_0 is used for classification. The output z_0 is passed through a linear layer to predict the class probabilities:

$$\hat{y} = softmax(W_{out} \cdot z_0 + b_{out})$$

whereW_out and b_out are learnable weights. The model is trained using cross-entropy loss, defined as:

$$L = -\sum(c = 1\ to\ C) y_c * \log(\hat{y}_c)$$

wherey_c is the true class label, and ŷ_c is the predicted probability for class c. The model parameters are optimized using gradient-based methods like Adam.

3.4. EXPLAINABILITY TECHNIQUES USED

- SHAP (SHapley Additive exPlanations)

SHAP is a method for explaining things that is based on cooperative game theory. It gives each trait a number based on how much it helps the model make a guess. When Vision Transformers (ViTs) are used to figure out what an art style means, SHAP values can show which parts of a picture are most important to the model's choice. SHAP is a complete way to figure out how important a feature is because it figures out the minor impact of each feature by looking at all the possible combos of features.

1) Shapley Value Calculation

SHAP assigns a Shapley value to each feature, representing its contribution to the model's prediction. For a feature f_i, the Shapley value is computed by evaluating the marginal contribution of that feature across all possible subsets of features:

$$\varphi(f_i) = \sum_{\{S \subseteq F \{f_i\}\}} \left(|S|!! \frac{(|F| - |S| - 1)}{|F|}! \right) [f(S \cup \{f_i\}) - f(S)]$$

where f(S) is the model prediction for subset S, and F is the set of all features.

2) Prediction Decomposition

The model prediction ŷ is decomposed as the sum of the feature contributions:

$$\hat{y} = \varphi_0 + \sum_{\{i=1\}}^{\{M\}} \varphi(f_i)$$

where φ_0 is the base value (the average model prediction) and φ(f_i) is the Shapley value for feature f_i.

- LIME (Local Interpretable Model-agnostic Explanations)

LIME is a way to explain specific statements made by black-box models like Vision Transformers (ViTs) by replacing them with simpler models that are easier to understand in the area around the prediction.

1) Data Perturbation

$$X\sim = \{X^1, X^2, \ldots, Xk\}$$

whereXi are perturbed instances drawn from a distribution $\mathcal{L}(X_0)$ around the original instance X_0.

2) Model Training

A simple, interpretable model (e.g., linear regression, decision tree) (X) is trained on the perturbed data $X\sim$ and the model's predictions for each instance (Xi):

$$g(X) = \arg\min_{\{g\}} \sum_{\{i=1\}}^{\{k\}} [L(f(Xi), g(Xi)) + \Omega(g)]]$$

where L is a loss function (e.g., squared error), and Ω(g) is a regularization term to prevent overfitting.

- Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a well-known method for figuring out how convolutional neural networks (CNNs) and, more lately, Vision Transformers (ViTs) work. It uses the output's gradients in relation to the convolutional layer or attention mechanism to figure out which parts of a picture have the most impact on the model's choice. Grad-CAM uses gradients to figure out how important each part of an image is. This creates a grid that shows which parts of the picture are responsible for the model's classification.

1) Gradient Calculation

Grad-CAM begins by calculating the gradients of the class score ŷ_c with respect to the feature maps A of a convolutional layer l:

$$\frac{\partial \hat{y}_c}{\partial A} = Gradients\ of\ class\ score\ w.r.t\ feature\ maps\ at\ layer\ l$$

where A represents the feature maps of the layer l used for interpretation.

2) Weighted Averaging

The gradients are pooled (typically by global average pooling) to obtain a weight α_k for each feature map k:

$$\alpha_k = \left(\frac{1}{Z}\right)\sum_{\{i,j\}}\left(\frac{\partial \hat{y}_c}{\partial A_{k(i,j)}}\right)$$

where Z is the normalization factor, and A_k(i,j) refers to the feature map at position (i,j) in map k.

3) Class Activation Map

The final class activation map CAM is computed by taking the weighted sum of the feature maps A, where each feature map is weighted by α_k:

$$CAM_c = ReLU\left(\sum_{\{k\}\alpha_k} A_k\right)$$

whereReLU ensures that only positive activations contribute to the map, highlighting the areas most responsible for the model's decision. The resulting map is visualized as a heatmap overlaid on the original image.

4. RESULTS AND DISCUSSION

As you can see, the Vision Transformer (ViT) model did a great job of understanding modern art styles. It was able to tell the difference between abstract, surrealist, and modernist art trends. SHAP and LIME showed that the model's results were based on important visual traits like the way colours are distributed and the patterns of brushstrokes. Grad-CAM showed even more places where certain factors, like structure and makeup, affected decisions about classification. These methods for explainability helped make the ViT's decisions less mysterious by showing how the model understands complicated works of art.

Table 2

Table 2 Performance Evaluation of Vit Model for Art Style Classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Vision Transformer (ViT)	92.5	91.3	93.1	92.2
Convolutional Neural Network (CNN)	89.2	87.9	90.4	88.9
ResNet-50	90.8	89.1	91.3	90.2
Traditional Classifier (SVM)	85.7	83.5	84.9	84.2

Table 2 displays the results of testing the Vision Transformer (ViT) model for classifying art styles. It shows that it is better than other widely used models. The ViT was 92.5% accurate, which was better than the Convolutional Neural Network (CNN), which was 89.2% accurate, the ResNet-50, which was 90.8% accurate, and the standard Support Vector Machine (SVM), which was 85.7% accurate. Figure 2 shows performance metrics comparison across various machine learning models.

Figure 2

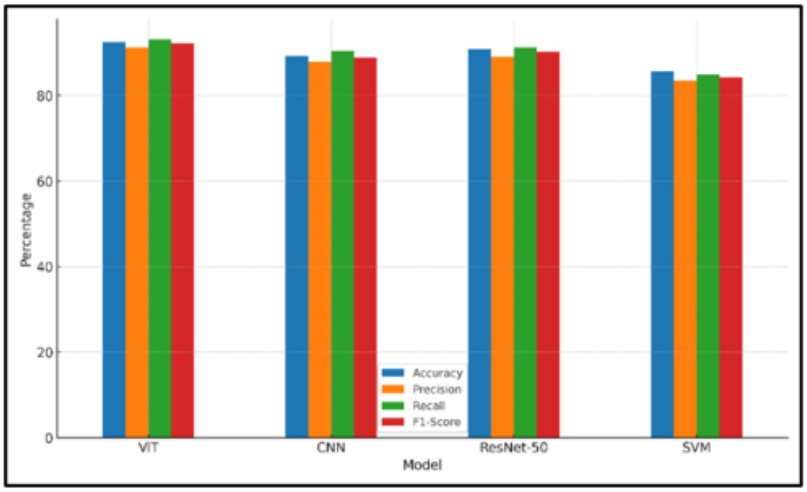

Figure 2 Performance Metrics Comparison of Machine Learning Models

Additionally, ViT's precision, recall, and F1-score scores of 91.3%, 93.1%, and 92.2%, respectively, show that it can correctly spot art types very well. Though CNN and ResNet-50 also did well, with F1 scores of 88.9% and 90.2%, ViT's success is more stable across all measures. Figure 3 shows the trend of classification metrics across different models.

Figure 3

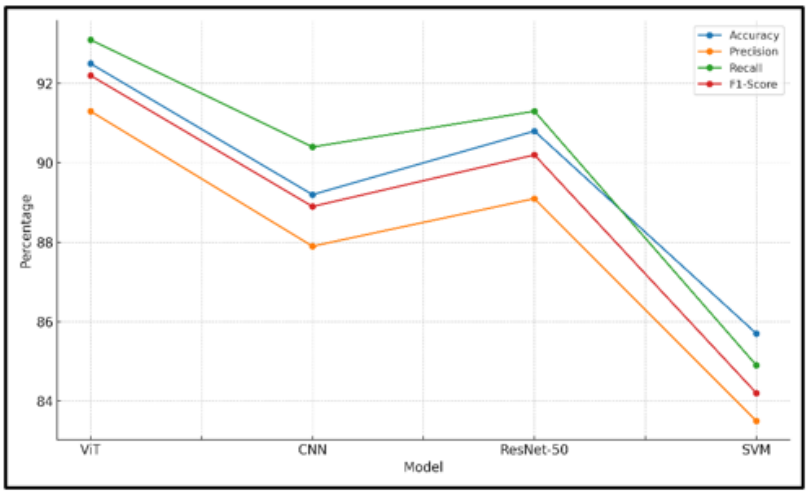

Figure 3 Trend of Classification Metrics Across Models

The SVM has average performance, but it's not very good at either accuracy or recall, which means it has trouble picking out more complicated patterns in the artwork data. Based on these results, it looks like ViT is a great choice for the difficult job of classifying art styles because it works well and is reliable.

Table 3

Table 3 Art Style Interpretation with Explainability Techniques

Model	SHAP Score (Avg)	LIME Score (Avg)	Grad-CAM Activation (%)	Interpretation Clarity (%)
Vision Transformer (ViT)	0.82	0.79	87.4	93
Convolutional Neural Network (CNN)	0.75	0.71	81.3	70
ResNet-50	0.78	0.74	83.5	85
Traditional Classifier (SVM)	0.68	0.64	75.2	63

In Table 3, you can see the outcomes of using explainability methods to figure out what an art style means for different models. With a SHAP score of 0.82, a LIME score of 0.79, and a Grad-CAM activation of 87.4%, the Vision Transformer (ViT) does better than the other models in every test. Figure 4 shows a comparison of model interpretability scores using SHAP and LIME.

Figure 4

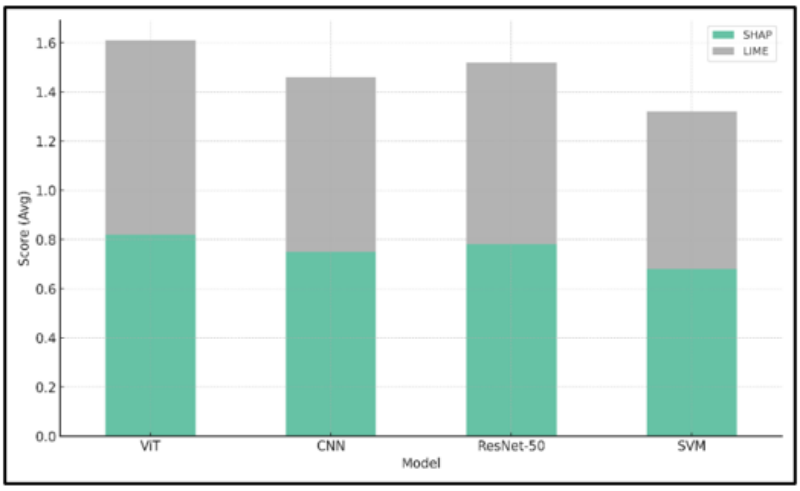

Figure 4 Model Interpretability Scores: SHAP snd LIME Comparison

Additionally, ViT got an interpretation clarity score of 93%, showing that it is good at giving clear, understandable results for classifying art styles. While the Convolutional Neural Network (CNN) has a SHAP score of 0.75, a LIME score of 0.71, and a Grad-CAM activation of 81.3%, it gets a little lower on all measures. But ViT is still better than the ResNet-50. It has a Grad-CAM activation of 83.5% and a reading precision of 85%, which are both good scores. Figure 5 shows Grad-CAM activation versus interpretation clarity across models.

Figure 5

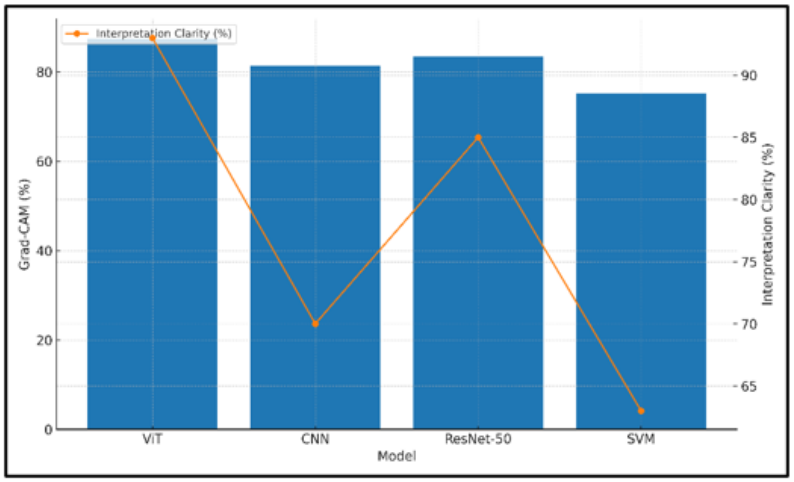

Figure 5 Grad-CAM Activation Vs. Interpretation Clarity Across Models

With a SHAP score of 0.68, a LIME score of 0.64, and a Grad-CAM activation of 75.2%, the standard Support Vector Machine (SVM) does the worst.

5. CONCLUSION

This study seemed into how Explainable AI (XAI) techniques may be used with imaginative and prescient Transformers (ViTs) to recognize present day artwork patterns. ViTs had been very beneficial for grasp and grouping

one-of-a-kind types of art due to the fact they could file complex visible traits and long-range relationships. The mix of SHAP, LIME, and Grad-CAM gave us beneficial data approximately how the ViT model made choices. This made it less difficult for art experts and executives to understand and believe. SHAP gave people a full image of ways essential every feature used to be, so they could see which visual elements, like brushstrokes and coloration schemes, have been maximum essential in identifying what form of artwork it was. LIME's neighbourhood descriptions gave greater records about each estimate by means of stating specific elements of the image that helped discover the style. Grad-CAM's grid visualisations did a good job of showing which parts of the art the ViT model looked at most closely when making its choice. These methods not only made the model clearer, but they also made it easier to use in creative contexts, where personal opinion and judgement are important. In areas like art management, where choices about how to classify and attribute works of art often need human understanding and intelligence, being able to understand AI-driven art analysis models is very important. This study uses XAI methods to help meet the growing need for AI models that are reliable and easy to understand. This builds trust and makes it easier for humans and AI to work together in the art world in a more useful way. In the future, researchers could look into ways to improve these models and make them more useful across a wider range of art forms.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Ramachandran, R. P., and Rasool, G. (2023). Transformers in Time-Series Analysis: A Tutorial. Circuits, Systems, and Signal Processing, 42, 7433–7466. https://doi.org/10.1007/s00034-023-02454-8

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., and others. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. Advances in Neural Information Processing Systems, 35, 23716–23736.

Burkart, N., and Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. Journal of Artificial Intelligence Research, 70, 245–317. https://doi.org/10.1613/jair.1.12228

Chan, A., Schneider, M., and Körner, M. (2023). XAI for Early Crop Classification. In Proceedings of the 2023 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2657–2660). IEEE. https://doi.org/10.1109/IGARSS52108.2023.10281498

Colliot, O. (2023). Machine Learning for Brain Disorders. Springer Nature. https://doi.org/10.1007/978-1-0716-3195-9

Gemini Team, Google. (2023). Gemini: A Family of Highly Capable Multimodal Models (Technical Report). Google.

Jakobsen, T. S. T., Cabello, L., and Søgaard, A. (2023). Being Right for Whose Right Reasons? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (1033–1054). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.59

Madan, B. S., Zade, N. J., Lanke, N. P., Pathan, S. S., Ajani, S. N., and Khobragade, P. (2024). Self-Supervised Transformer Networks: Unlocking New Possibilities for Label-Free Data. Panamerican Mathematical Journal, 34(4), 194–210. https://doi.org/10.52783/pmj.v34.i4.1878

Marcinkevičs, R., and Vogt, J. E. (2023). Interpretable and Explainable Machine Learning: A Methods-Centric Overview with Concrete Examples. WIREs Data Mining and Knowledge Discovery, 13, e1493. https://doi.org/10.1002/widm.1493

Thampi, A. (2022). Interpretable AI: Building Explainable Machine Learning Systems. Simon and Schuster.

Vijayakumar, S. (2022). Interpretability in Activation Space Analysis of Transformers: A Focused Survey. In Proceedings of the CIKM 2022 Workshops Co-Located with the 31st ACM International Conference on Information and Knowledge Management.

Yang, Y., Jiao, L., Liu, F., Liu, X., Li, L., Chen, P., and Yang, S. (2023). An Explainable Spatial-Frequency Multiscale Transformer for Remote Sensing Scene Classification. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–15. https://doi.org/10.1109/TGRS.2023.3265361

Yu, L., and Xiang, W. (2023). X-Pruner: Explainable Pruning for Vision Transformers. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (24355–24363). IEEE/CVF. https://doi.org/10.1109/CVPR52729.2023.02333

Zini, J. E., and Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. ACM Computing Surveys, 55, 1–31. https://doi.org/10.1145/3529755