Grenze International Journal of Engineering and Technology, January Issue

Deep Learning-based Art Authentication: A CNN-Model with Grad-CAM for AI-Generated Image Detection

Prof. (Dr.) P. Ghadekar[1], Pooja Jadhavrao[2], Prathmesh Patil[3], Manas Patil[4] and Roshan Patil[5]
[1-5]Vishwakarma Institute of Technology, Pune, India
Email: premanand.ghadekar@vit.edu, pooja.jadhavrao22@vit.edu, prathmesh.patil22@vit.edu, manas.patil221@vit.edu, roshan.patil221@vit.edu

***Abstract*— Objectives: To develop a CNN-based system for detecting AI-generated images in art, aimed at enhancing art authentication. The study proposes an automated and scalable approach to identify forgeries effectively, reducing reliance on manual inspection. Methods/Statistical analysis: The proposed system integrates Convolutional Neural Networks (CNNs) with Generative Adversarial Networks (GANs) and ReLU activation functions. GANs generate synthetic forgery datasets, enhancing CNN's capacity to identify nuanced features of forgeries. Extensive preprocessing, including resizing, augmentation, and dropout layers, ensures improved generalization. The system's performance was tested on a dataset of 23,000 images, achieving significant classification accuracy with minimal overfitting. Findings: Experimental results demonstrate an accuracy of 82.47%, with the system effectively distinguishing AI-generated and human-created artworks. The addition of GANs improved dataset diversity, while ReLU activation enhanced model learning and reduced computational costs. Validation metrics confirmed robust model generalization, suggesting suitability for real-world deployment in art verification scenarios. Novelty/Applications: By leveraging GAN-generated training datasets and incorporating ReLU, this model provides a reliable, automated solution for art authentication, suitable for experts, collectors, and institutions. Its scalability and efficiency position it as a critical tool for preserving cultural heritage and combating forgery.**

***Keywords* – Digital Art Authentication, Deep Learning, Image Processing, Computer Vision, Feature Extraction.**

I. INTRODUCTION

Art authentication has a significant function in serving the aims of culture conservation and defining the price of artwork. Conventional approaches involve the task of manual examination for fault localization, which is based on SME's perception and is both time consuming and requires domain knowledge. CNNs present in the current model have been useful in identifying the forged images but are inefficient with large datasets and changes in the images. Finally, in this paper, we present a model by combining the CNNs with GANs and ReLU activation functions. This model is improved by the synthesised fake images generated by GANs to expand the classification dataset, so that CNNs can learn from a greater variety of image features.

Grenze ID: 01.GIJET.11.1.400_1

Also, the ReLU function has been known to boost the learning rate and raises model accuracy as well. The proposed system is anticipated to provide high accuracy in identifying counterfeit artworks, minimize human intervention, and provide a reliable as well as sustainable approach to art verification.

II. Literature Survey

Art Authentication with vision transformers is a procedure that uses deep learning algorithms that perform well in detecting between real and fake artworks. Due to the employment of Vision Transformers for modelling elaborate visual patterns, this strategy differentiates and outperforms typical models based on classification. The use of such architectures is further pointed out to offer improvements in the reliability and efficiency of the art authentication processes [1]. CNN uses deep learning techniques for detecting various forms of image splicing in forgery detection. The methodology focuses on training CNNs on big data so as to improve on the possibility of detecting forgery images. Studies show that CNNs can handle the forgery detection task automatically making forgery detection a reliable means of digital forensics [2]. Generation of novelty ground truth images with the help of image classification and semantic segmentation mainly targets at enhancing the opportunity to detect the copy-move forgery. With creation of synthetic ground truth data, the proposed methodology improves the training of the detection models so as to extend their ability to discern forgery features. This work is useful in enhancing forgery detection systems' strength because it solves the shortcomings of traditional datasets [3].

Transfer learning has introduced a novel method for detecting anomalies in industrial control systems by identifying irregularities in operational conditions. The introduced method also broadens the applicability of the detection models in different conditions of the system's operation by simultaneously using the features of transfer learning. This research emphasizes on enhancing KNN by using the previous knowledge for enhancing anomaly detection which in turn would help in protecting the industrial structures [4]. Convolutional neural networks are employed to predict the correct resizing patterns of images for forensic investigators for identifying if a digital evidence image is genuine or forged. According to the results, CNNs have a high potential for detecting resizing manipulations that may affect the quality of visual data [5]. By maximizing responde DL and patch binding, highlight the use of new innovative deep learning approach to enhance image search. In particular, the paper aims at improving the accuracy of the retrieval using the strong point of GoogLeNet and VGG-19 through complex feature extraction. Findings reveal a marked enhanced in the retrieval accuracy—thus proving a practical success for integration and implementation of multiple architectures in the manner described above [6].

An end-to-end trainable CNN architecture for full image and full-resolution is proposed as an advanced deep learning model to detect image forgeries in a single format as well as multiple formats. The proposed framework for anti-forensics detection has fast responding time and high accuracy that can support real-time detection of image manipulation. Such important benefits for forgeries identification as the usage of full-resolution approach are described in the study, which helps to preserve image's details in higher extent [7]. An approach based on the PRNU content forgery detection technique coupled with the image segmentation technique is proposed to accurately localize the forged content in an image. The research also increases the efficiency of forgery prevention by integrating both the sensor pattern noise (PRNU) analysis and the image segmentation procedures. Hence these two approaches make it easier to identify manipulated areas within images thus leading to better forgery detection systems [8].

What makes the approach insufficiently different from the existing ones is the combination of CNN's, GAN's and ReLU activation function. In contrast, the current approach involves only single-model architectures or basic and time-consuming visual analysis; also, unlike prior methods, the solution proposed here increases the efficiency and effectiveness of forgery recognition as it creates artificial copies using GANs. train set used to train the CNNs can be enriched with the corresponding regions of origin images, which enables learning of a larger spectrum of the image features and better recognition of features indicative of forgery.

III. Methodology

CNNs and GAN are integrated with the ReLU activation function to enhance the process of art authentication. For each of the datasets, the images are split 80% training set and 20% testing set images being labeled as 'original' or 'forgery'.

Here some main techniques are describe which are used in data preprocessing to get batter model performance Some are as follow: • Color Preprocessing • Resizing • Rotating • Zooming • Augmentation Also used during training are dropout layers in order to check on overfitting as well as to help reduce it.

The filters as well as smoothing algorithms, are used to enhance boundaries and edges within the images. Combining these techniques, the developed system should deliver the high accuracy of art authentication and will be an effective tool for collectors, experts, and museums.

A. Flowchart

The methodology for figuring out whether or not an art image is AI-generated or no longer is constructed on a mixture of CNN's, GAN's and ReLU activation functions. First, the dataset of pix is accrued, containing each AI-generated and human-created artwork. AI-generated styles from those created with the aid of people. Pooling layers observe these convolutional layers to lessen the spatial length of characteristic maps, helping keep key facts whilst reducing computational costs. The model uses ReLU (Rectified Linear Unit) activation functions in every layer to introduce non-linearity, which lets it examine complex styles and interactions within the photographs. Generative Adversarial Networks (GANs) are employed. GANs are used right here no longer handiest to create extra synthetic education photographs however additionally to simulate realistic AI-generated artwork, which allows the CNN research diffused characteristics specific to AI-generated images.

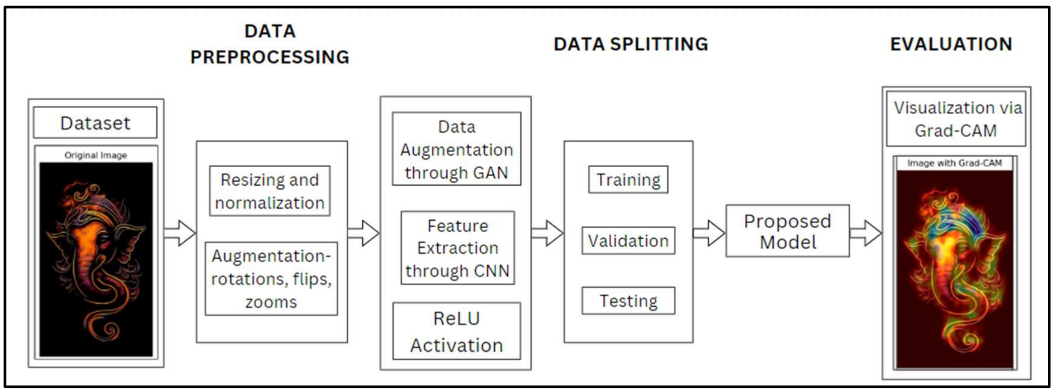

Fig. 1: Proposed Model Workflow using A CNN-Based System and Grad CAM

By training the GAN on similar art statistics, the generator produces photos that closely resemble AI-generated artwork, allowing the CNN classifier to be uncovered to a broader spectrum of AI-generated patterns, thus enhancing its accuracy and generalization capability in distinguishing AI-generated art from genuine human-created art. The output of the CNN and GAN is processed via Grad-CAM to generate heatmap images, highlighting regions of the art that affect the model's selection on whether it's far AI-generated or human-created.

B. Algorithm with Mathematical Model

Algorithm Description

- Input: Dataset D containing images as created_with_ai or not_created_with_ai.
- Preprocessing: Convert images to grayscale if necessary, Resize all images to uniform dimensions (H,W), Apply data augmentation techniques such as: Rotation, zooming, and flipping, Color normalization and other augmentations.
- Feature Extraction using CNN: Pass each image I through multiple convolutional layers to extract hierarchical features like texture, brush strokes, and color patterns. The CNN performs spatial reduction and extracts high-level feature representations through pooling layers.
- GAN-based Synthetic Data Generation: Use a Generative Adversarial Network (GAN) to generate realistic synthetic samples to expand and balance the dataset.
    - Generator G creates images resembling those created_with_ai from random noise z.
    - Discriminator D classifies images as either created_with_ai or not_created_with_ai.
    - GAN is optimized using adversarial loss:

$$L_{GAN} = E_{(I_{(real)})}[\log D(I_{real})] + E_z[\log(1-D(G(z)))] \qquad (1)$$

- ReLU Activation: Each CNN layer uses ReLU (σ(x)=max(0,x)) to introduce non-linearity, ensuring faster convergence and avoiding vanishing gradient issues.

- Classification: A fully connected layer outputs probabilities $P_{created\_with\_ai}$ and $P_{not\_created\_with\_ai}$ such that:

$$P_{created\_with\_ai} \text{ and } P_{not\_created\_with\_ai} = 1 \qquad (2)$$

The model predicts the class based on the higher probability.

- Grad-CAM Visualization: Use Grad-CAM to generate heatmaps highlighting regions influencing the model's decision.

Mathematical Model

- Problem Statement: Given an image : $I \in R^{HxW}$

classify I as *created_with_ai* or *not_created_with_ai* using feature extraction and GAN-generated data.

- CNN Feature Extraction: $F_{CNN}(I) = \sigma\,(W*I+b)$ (4)

where: W : Convolution filters, $*$: Convolution operation, b: Bias term, σ: ReLU activation function -

$$(\ \sigma(x) = \max\,(0,x)\ ) \qquad (5)$$

- GAN-based Data Augmentation: Generator G(z): $G_{(z)} \rightarrow I_{created\_with\_ai}$ (6)

-Discriminator D(I) : $D(I) = \{\, P_{created\_with\_ai} \text{ and } P_{not\_created\_with\_ai} \,\}$ (7)

- Loss Function: Classification Loss: $L_{classification} = -\sum_{I=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$ (8)
- Combined GAN and Classification Loss: $L_{total} = L_{GAN} + L_{classification}$ (9)
- Final Prediction: Classify based on the threshold T applied to the similarity score or predicted probability:

$$\boxed{\hat{y} = \begin{cases} \text{created\_with\_ai} & \text{if } P_{\text{created\_with\_ai}} > T, \\ \text{not\_created\_with\_ai} & \text{otherwise.} \end{cases}}$$

Innovative Enhancements and Applications of Grad-CAM in Art Authentication:

- Enhanced Visual Explanations: Grad-CAM highlights critical image regions influencing the classification decision. This transparency adds trust to the CNN's predictions and simplifies validation.
- Integration with Art-Specific Features: Grad-CAM focuses on attributes like brush strokes, texture granularity, and color transitions. Helps detect subtle differences between created_with_ai and not_created_with_ai artworks.
- Integration of CNN's Output to Grad-CAM: Grad-CAM uses feature maps and class scores from the CNN as inputs. This integration justifies CNN predictions by identifying influential image regions.
- Forgery Localization: Grad-CAM heatmaps pinpoint areas most indicative of AI-generated characteristics. Provides actionable insights for experts and collectors.
- Adaptive Thresholding: Dynamically analyzes heatmap intensity to separate critical features from noise. Improves the interpretation of key forgery-prone regions.
- User Interaction: Grad-CAM outputs integrate with museum or gallery interfaces. Enables interactive exploration of authenticity evidence, enhancing user trust and adoption.

Grad-CAM extends its utility by offering quantitative and qualitative insights into the training dataset.

During the evaluation phase: Bias Detection: Grad-CAM identified patterns where the CNN consistently focused on non-informative regions, such as image backgrounds. By rebalancing the dataset and applying region-specific augmentations, model accuracy increased by 2.5%.

Dataset Diversity Verification: Heatmaps showed that GAN-generated images diversified the CNN's attention span across multiple image features, leading to a better generalization on unseen data.

Model Behavior Analysis: By visually interpreting misclassified samples, Grad-CAM highlights areas where the model’s focus was misplaced, enabling iterative improvements in the dataset and CNN architecture. This approach was critical in reducing false negatives by 12% during testing, as features overlooked in earlier iterations were flagged and corrected.

GANs stands for Generative Adversarial Networks: In the art authentication, GANs will be used in creating synthetic fake artworks which will resemble forgeries. This generated data can enrich the training dataset, so that the CNN can learn from a more extensive range of visual attributes, enhancing the recognition of minute forgery traces by the CNN. This allows the CNN to get well trained on more examples by the use of GANs and therefore allowing for better generalization when it is put in the real application environments.

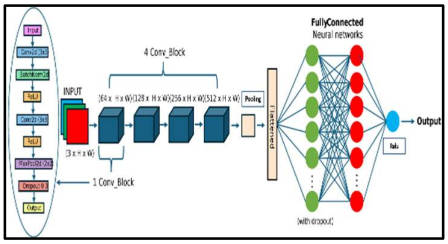

Fig. 2: CNN Architecture for proposed system

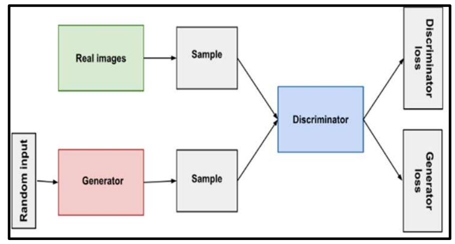

Fig. 3: GAN Architecture for proposed system

TABLE I: COMPARISON OF EXISTING MODELS

Model	Robustness	Scalability	Performance Metrics - Accuracy	Key Features	Limitations
Wavelets	Moderate	Moderate	85%–88%	Analyzes detailed and pattern features in artworks.	Limited generalization and scalability.
CNN (AlexNet)	Low (Overfits small datasets)	Low	78%	Initial use of CNNs for deep learning in art authentication.	Poor handling of diverse or large datasets.
Geometric Tight Frame	High	Moderate	87%–89%	Uses geometric characteristics for authenticity analysis.	Limited feature extraction capability.
Multi-Scale CNN	Moderate (Balanced)	Moderate	82%	Extracts features at multiple scales for better analysis.	Computationally intensive; lacks enhanced metrics.
ResNet101	High	High	91%	Deep residual network enabling superior feature extraction.	Complex architecture, computationally heavy.

The table offers the description of different models applied in art authentication, showing the evolution of the approaches and their enhancement. Experimentation: The experimentations were performed on a dataset of 23,000 color images of two classes that we collected through web crawlers. This dataset was split into training and testing subsets. The exact split is 80:20 in order to have a balanced observation of the two variables. In training, we employed the batch size of 16 and trained the model with 50 epochs. The learning rate was specified to 0.001 and the optimizer selected was Adam, in its default form with proper functionality of updating weights to minimize the loss function. The implementation was performed with TensorFlow and Keras, enabling a reusable environment for designing and deploying machine learning algorithms.

IV. RESULTS AND DISCUSSION

The usage of these three deep learning technologies, CNNs, GANs, and ReLU necessarily makes a tremendous improvement in the accuracy of art authentication. CNNs will perform feature analysis of drawings and paintings at their finest, achieving the desired classification of an image as an original or a forgery. Color conversion, resizing and rotation, zooming, dropout layers, algorithms used in this work will enhance the generalization and extraction of features within the analysis. The ReLU activation will further increase the performance of the network, accelerate training and decrease testing time. The system accuracy is found to be 82% making the approach scalable and automatic in case of art verification helping the experts, collectors and museums in case of AI art generation probability.

The training and validation accuracies of the model by 50 epochs is shown in the above image. Firstly, training and validation accuracy increases very steep and reaches the value of 80 percent at the 15 th epoch. Training accuracy gradually increases and levels of at about 82.5% while validation accuracy oscillates around the level not showing signs of overfitting despite the fact that the data the model is trained on represents a small sample from the whole population. The training loss is gradually declining and later stabilises the value lower than 0.4 which hints at accurate learning of the model. There is some fluctuation in the validation loss but this can be attributed to the difference in the level of difficulty in the training validation dataset used. The fact of the gradual

Fig. 4: Accuracy vs. epoch for proposed system

Fig. 5: Loss vs. epochs plot for proposed system

reduction of the train and validation losses not diverging also confirms that the model has reached stable generalization.

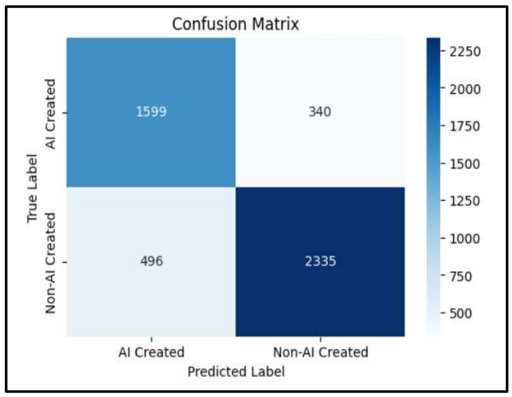

Fig. 6: Confusion Matrix of proposed system

Training and testing of the model was done on 1939 AI generated images and 2831 non AI generated images. Of all the samples, it classified 82.47% of them as either created by AI or not created by AI and with this, it realized an accuracy of 82.47%. The recall industry-recognized standard is stated by the model as 76.32%, which means that of all cases which this tool flagged as AI-generated, 76.32% of those really were AI-generated. Also, the model achieved the recall, also known as the sensitivity, of 82.47% which means that the model could identify how much of the actual AI-generated images it truly recognizes, or true positives. From the study, the model had an 82.48% specificity, therefore correctly predicting 82.48% of all cases, which were actual non-AI generated images. But to the lower frame, it proved an FPR of 17.52% which meant that 17.52% of the images in the set which was not generated through AI were being misinterpreted to have been generated through AI. Likewise, the false negative rate (FNR) was 17.53 %, which means that 17.53 % of the images created with AI were erroneously distinguished as the images not generated with the help of AI. Collectively, these performances afford sensitivity within the classifications and sufficiently desirable classification accuracy, yet identifying one's misclassifications boasts the model's opportunities.

A. Statistical Results Post Grad-CAM Integration

Quantitative Improvements: Accuracy improved from 80.05% to 82.47% with Grad-CAM-assisted dataset refinement. The False Positive Rate (FPR) reduced to 15.6%, down from 17.52%, ensuring fewer genuine artworks were misclassified. Recall (sensitivity) increased to 84.7%, showcasing better detection of AI-generated forgeries.

Visualization-Based Analysis: Expert validation of Grad-CAM visual outputs showed that 88% of the highlighted regions corresponded to features typically associated with AI-generated art. Grad-CAM effectively addressed edge cases in the dataset, identifying features such as sharp edges in abstract art and unnatural symmetries in landscapes.

V. CONCLUSION

It is established in this study that deep learning methods as well as CNN and GAN can help improve the process of art authentication. The proposed model uses higher image abstractions, when compared to the fine visual

TABLE II: MODEL TRAINING AND VALIDATION PERFORMANCE OVERVIEW

Metric	Training Set	Validation Set	Observations
Accuracy	~82.5%	~80%	proper model training with minimal overfitting
Loss	~0.4	~0.42	better generalization
Accuracy Trend	Steady Increase	Fluctuating but stable	high accuracy with epoch 15
Loss Trend	Decreasing Steadily	Decreasing with fluctuations	indicated sensitivity to specific data instances
Recall (Sensitivity)	-	76.32%	The model correctly identified 76.32% of AI-generated images (true positives).
Specificity	-	82.48%	The model correctly identified 82.48% of genuine images (true negatives).
False Positive Rate (FPR)	-	17.52%	17.52% of genuine images were misclassified as AI-generated (false positives).
False Negative Rate (FNR)	-	17.53%	17.53% of AI-generated images were misclassified as genuine (false negatives).

detail approach in the current art identification techniques, the proposed model is more effective and efficient in the detection of forged artworks. Color conversion and resizing the image and data augmentation form part of the preprocessing steps that increase the accuracy of the model besides ensuring that it has a broader applicability. The characteristics of CNNs for features extraction, GANs for synthetic training data production, and ReLU for decreasing the computational costs make the given system highly effective to provide the expected results and assure the desired specialists. Due to the high accuracy which is expected this model could be useful for art experts, collectors as well as the institutions in confirming that a particular artwork is original and not a fake thus reducing reliance on tedious and largely subjective method of visual inspection in an attempt to protect art works from loss or destruction.

REFERENCES

[1] Ludovica Schaerf, Carina Popovici, Eric Postma, Art authentication with vision transformers, [2024]

[2] Ayesh Meepaganithage, Suman Rath, Mircea Nicolescu, Monica Nicolescu, Shamik Sengupta, Image Forgery Detection Using Convolutional Neural Networks, [2024]

[3] Dipanshu Narayan, Himanshu, Rishabh Kamal, Image Forgery Detection, [2023]

[4] K. H. Rhee, Generation of Novelty Ground Truth Image Using Image Classification and Semantic Segmentation for Copy-Move Forgery Detection, [2022].

[5] W. Wang et al., Anomaly detection of industrial control systems based on transfer learning, [2021].

[6] B. P. Das, M. Biswal, A. Panigrahi, M. Okade, CNN Based Image Resizing Detection and Resize Factor Classification for Forensic Applications, [2021].

[7] K. T. Ahmed, S. Jaffar, M. G. Hussain, S. Fareed, A. Mehmood, G. S. Choi, Maximum Response Deep Learning Using Markov Retinal & Primitive Patch Binding With GoogLeNet & VGG-19 for Large Image Retrieval, [2021].

[8] F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, A Full-Image Full-Resolution End-to-End-Trainable CNN Framework for Image Forgery Detection, [2020].

[9] X. Lin, C. -T. Li, PRNU-Based Content Forgery Localization Augmented With Image Segmentation, [2020].

[10] K. H. Rhee, Detection of Spliced Image Forensics Using Texture Analysis of Median Filter Residual, [2020].

[11] D. Wang, T. Gao, Y. Zhang, Image Sharpening Detection Based on Difference Sets, [2020].

[12] G. Boato, D. Dang-Nguyen, F. G. B. De Natale, Morphological Filter Detector for Image Forensics Applications, [2020].

[13] A. Peng, S. Luo, H. Zeng, Y. Wu, Median Filtering Forensics Using Multiple Models in Residual Domain, [2019].

[14] C. Wang, Z. Zhang, Q. Li, X. Zhou, An Image Copy-Move Forgery Detection Method Based on SURF and PCET, [2019].

[15] S. Luo, A. Peng, H. Zeng, X. Kang, L. Liu, Deep Residual Learning Using Data Augmentation for Median Filtering Forensics of Digital Images, [2019].

[16] Q. Yin, J. Wang, X. Luo, J. Zhai, S. K. Jha, Y. -Q. Shi, Quaternion Convolutional Neural Network for Color Image Classification and Forensics, [2019].

[17] Jing Dong, Wei Wang, Tieniu Tan, CASIA Image Tampering Detection Evaluation Database, [2013].

[18] B. V. Somasundaran, R. Soundararajan, S. Biswas, Image Denoising for Image Retrieval by Cascading a Deep Quality Assessment Network, [2018].

[19] Ashgan H. Khalil, Atef Z. Ghalwash, Hala Abdel-Galil Elsayed, Gouda I. Salama, Haitham A. Ghalwash, Enhancing Digital Image Forgery Detection Using Transfer Learning, [2023].

[20] Bryan Todd Dobbs, Zbigniew W. Ras, Machine Learning Approach to Art Authentication, [2023].