Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Application of Two-Stream Late Fusion on EfficientNetV2 based on Transfer Learning to classify AI-generated paintings

Muhammad Kevin Rinaldi [1)], Ernawati [2)] , Desi Andreswari [3)] , Julia Purnama Sari [4)]
[1,2,3,4)] Informatics Study Program, Faculty of Engineering, University of Bengkulu, Bengkulu, Indonesia
[1)]mkevinrinaldi@gmail.com, [2)]ernawati@unib.ac.id, [3)]desi.andreswari@unib.ac.id,
[4)]juliapurnamasari@unib.ac.id

Submitted :Jan 14, 2026 | **Accepted** : Feb 4, 2026 | **Published** : April 2, 2026

Abstract: The rapid advancement of generative artificial intelligence (AI) has made synthetic digital paintings increasingly difficult to distinguish from human-made artworks, raising concerns regarding authenticity, copyright protection, and digital forensics. The main objective of this research is to develop a reliable and interpretable framework for distinguishing AI-generated paintings from human-created artworks by integrating visual and noise-based features. To address the limitations of conventional single-stream CNN models, this study proposes a Two-Stream Network with a Late Fusion strategy, combining a visual stream based on EfficientNetV2-S and a noise stream based on Xception with Spatial Rich Models (SRM).The proposed architecture processes semantic visual features and residual noise characteristics independently, followed by weighted decision-level fusion with a ratio of 0.7:0.3. Experiments were conducted using the AI-Artwork public dataset from Kaggle, consisting of 15,000 images with a data split of 64% training, 16% validation, and 20% testing. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, ensuring a comprehensive assessment beyond accuracy alone. The results demonstrate that the proposed method achieves 98% accuracy, 98% precision, a 99% F1-score, and high discriminative capability compared to single-stream baselines. Model interpretability was analyzed using Grad-CAM to examine the contribution of each stream. Despite promising results, this study is limited by evaluation on a single dataset and static fusion weights, which may affect generalization to unseen generative models. Future work includes cross-dataset evaluation, adaptive fusion strategies, and exploration of lightweight architectures. Practically, this approach has potential applications in digital art authentication, forensic analysis, and content moderation systems, as well as supporting emerging policies for AI-generated content regulation and copyright protection.

Keywords: AI-Generated Art; EfficientNetV2; Late Fusion; Two-Stream Network., Grad-CAM

INTRODUCTION

Rapid technological developments have brought about major changes in various fields, one of which is digital art, which has undergone a transformation due to the emergence of Artificial Intelligence technology. This phenomenon is marked by the emergence of AI capable of creating visual artworks (Anggraini et al., 2024). However, its development has sparked controversy regarding authenticity in digital art, given that this technology can produce visuals that are almost identical to human works (Aris et al., 2023). Alongside advances in digital art, there are challenges in the realm of digital art security. This is evident in the ease of access to generative AI, which has triggered a surge in synthetic content that has the potential to violate copyright and facilitate visual disinformation. Furthermore, AI training is carried out without explicit permission from the original owner of the work, which can constitute plagiarism of artistic works. Therefore, the development of methods for detecting AI-processed artwork has become increasingly crucial to maintain the integrity of artwork information (Li & Stamp, 2025). Manual detection methods are no longer effective in dealing with the sophistication of generative AI, as its

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

developments are becoming increasingly difficult to detect. Thus, an automated and accurate computational approach is needed (Mahara & Rishe, 2025).

In response to this challenge, a Deep Learning-based approach, specifically the Convolutional Neural Networks (CNN) architecture, has become a solution for recognizing complex image patterns (Alzubaidi et al., 2021). This has been proven in various case studies, particularly in AI painting detection. Previous research using various modern CNN architectures has shown high performance in detecting AI paintings (Bianco et al., 2023). Similarly, Deepfake detection in facial images achieved 91% accuracy using CNN architectures such as EfficientNet (Adventino Gulo et al., 2025). Success in the domain of deepfakes on faces is relevant to paintings because both are produced through similar generative AI mechanisms, namely learning statistical data distributions and reconstructing images through a sampling process (Corvie et al., 2022). Although the CNN method is effective in AI detection, previous studies have shown that AI detection has additional challenges due to variations in painting artistic styles (Subkhi et al., 2023). Furthermore, AI generative images show that synthetic images still leave artifacts and noise in the residual noise domain, which can be used as clues to distinguish between real and artificial images (Corvie et al., 2022). Therefore, due to problems in previous studies that still rely on visual content, this study applies the use of SRM, which has the potential to reveal noise in generative AI. This method is applied because AI-generated images, although visually realistic, exhibit statistical differences in residual noise. Although many studies have examined the use of CNNs in synthetic image detection, deepfake detection, noise analysis through Spatial Rich Models (SRM), and the application of model interpretability techniques. However, to date, there has been no research that explicitly integrates visual features and noise in the Two Stream Network architecture with the Late Fusion mechanism for AI painting classification, while also being validated interpretatively using Explainable AI. Most studies still focus on the domain of faces or general photographic images and tend to use only one type of feature, without combining the two (Atrey et al., 2010). In this study, the Two Stream Networks method was used by combining two different input streams to enrich the features analyzed (Simonyan, 2014). However, this study has several limitations, including the use of a dataset that is still limited to certain types and styles of paintings, so that its ability to generalize to variations in painting styles and the latest AI generative models has not been fully tested. In addition, this study is still limited to binary classification and depends on image resolution quality, so that performance may decline in images that have undergone heavy compression.

Based on this background, the main objective of this study is to develop and evaluate a CNN-based Two-Stream Network architecture with a Late Fusion mechanism for detecting AI-generated paintings, as well as to validate the model's decisions using Explainable AI techniques. The scope of the study focuses on binary classification between human-made paintings and AI-generated paintings, analyzing the contribution of visual features and residual noise to the performance and interpretability of the model. This research contributes conceptually by expanding the paradigm of synthetic image detection to the realm of painting, and proves that the integration of visual and noise streams in the Two-Stream Late Fusion framework improves classification reliability and model transparency through Grad-CAM visualization.

LITERATURE REVIEW

Along with the rapid development of technology, there has also been an improvement in AI in the process of synthesizing images such as paintings with high quality. This improvement in AI capabilities has blurred the clear boundaries between human-made art and art produced by AI. Now AI can imitate artistic styles, compositions, and textures that resemble original human-made paintings. This phenomenon has triggered research related to the detection of painting forgeries using AI. In previous studies, the application of Deep Learning methods was widely used because it can perform image detection, particularly the Convolutional Neural Networks (CNN) architecture, which has become the most relevant method for recognizing complex image patterns (Alzubaidi et al., 2021).

The effectiveness of this method has been proven in various case studies in AI detection. One study that used CNN to detect Deepfakes in facial images achieved an accuracy of 91% by applying the CNN architecture (Adventino Gulo et al., 2025). Although many early studies focused on facial images, these findings remain relevant to the domain of paintings because both faces and paintings are generated by similar generative models and both leave noise patterns. In painting image detection, the use of CNN architectures such as VGG19 can achieve an accuracy of 96.88% in only the second epoch. Although this figure is high, the study reports that there are indications of overfitting because the VGG19 model has a heavy structure and is highly dependent on sensitive visual feature extraction (Subkhi et al., 2023). Similar findings were reported by Li and Stamp, where the use of CNN on a painting dataset achieved a high accuracy of 97.58%, but performance decreased significantly to 82.08% in the multi-class classification task (Li & Stamp, 2025). Therefore, an architecture capable of overcoming heavy computation and slow training is needed, and the EfficientNetV2 architecture was chosen for its efficiency. This can be proven in Vivekananda's research on strengthening digital security through deepfake detection. This study recorded high performance results in deepfake detection, with the EfficientNetV2 model achieving high accuracy

*name of corresponding author

of 99% with minimal error (Vivekananda et al., 2025). Furthermore, a Transfer Learning strategy with Fine-Tuning was applied as a solution to overcome the problems of overfitting and large computations (Cetinic et al., 2018). Research using the Fine-tuning strategy explains that the features learned in the early layers tend to be general, so the use of fine-tuning allows pre-trained models to be adapted to new tasks with high efficiency (Yosinski et al., 2014).

AI generative models, despite being capable of producing images with high visual quality, have the disadvantage of leaving residual noise traces. A number of studies show that generated images have structured noise patterns and pixel correlations that differ from natural images, originating from the neural network-based generation process (Zhang et al., 2025). Therefore, this study applies a method of analysis through noise in images because the noise left behind by AI can be used as forensic evidence to distinguish between artificial and original images (Corve et al., 2023). To extract noise traces, the use of a high-pass filter is necessary because it can remove visual information and highlight fine noise components in the image (Zhang et al., 2025). Therefore, this study applies a technique that uses a high-pass filter, namely Spatial Rich Models (SRM), which can extract residual noise components by suppressing visual content information and highlighting noise patterns and local dependencies between pixels (Fridrich & Kodovský, n.d.).

However, previous research results still have a number of methodological weaknesses. Most CNN-based approaches in AI painting detection are highly dependent on visual features, thus potentially experiencing overfitting and performance degradation when faced with artistic style variations or multi-class scenarios. In fact, visual analysis often has limitations, especially when dealing with subtle areas or noise in images that are difficult to see. On the other hand, although noise analysis and Spatial Rich Models (SRM) have proven effective in the Deepfake domain, their application to painting images is still limited and generally does not stand alone without direct integration with CNN-based visual features in a single framework. The single-stream approach has limitations in comprehensively representing the characteristics of AI paintings. Visual-based models alone tend to be sensitive to variations in artistic style and texture (Li & Stamp, 2025). Meanwhile, noise models have the potential to ignore the semantic context and compositional structure of paintings. Therefore, this study applies noise analysis using a hybrid approach with the Two Stream Networks method and Late Fusion mechanism. This has been proven in research that combines two different input streams to enrich the analyzed features (Simonyan, 2014). Several studies have proposed the Two-Stream approach to detect image manipulation, particularly in the domain of faces, by combining visual content and noise (Morariu & Davis, n.d.). However, this approach has not been specifically applied to the classification of AI-generated paintings, nor has it examined the contribution of each stream in the context of painting, which has more complex variations in style, texture, and composition.

In addition, most previous studies have not systematically utilized Explainable AI to verify that the model's decisions are truly based on noise traces and visual forms (Castellano et al., 2024). Given that models are susceptible to the Hans Clever Effect, a phenomenon where models achieve high accuracy by exploiting false correlations or irrelevant data features, instead of learning according to context-appropriate features, the model seeks shortcuts for learning. To address this issue, the Gradient-weighted Class Activation Mapping (Grad-CAM) technique was applied as a visualization method that highlights important areas in the image that form the basis of the model's prediction (Castellano et al., 2024).

Thus, it can be concluded that there is a clear research gap, namely the absence of research that integrates visual features and noise in the Two-Stream Network architecture with the Late Fusion mechanism for classifying AI-generated paintings as a more reliable alternative compared to the single-stream approach that relies on only one type of feature. Thus, the contribution of this research is to show that the integration of two complementary streams between visual features and noise produces better performance and robustness compared to single-stream architectures, while also providing interpretability validation to ensure that model decisions are based on relevant visual and forensic traces.

METHOD

This research was conducted experimentally using quantitative methods to evaluate the performance of the Two Stream Networm architecture in classifying human-made paintings and Artificial Intelligence (AI). The Two Stream Late Fusion method processes images through two independent streams: the visual stream (RGB Stream) to capture visual image features and the noise stream (Noise Stream) to detect noise traces. These two streams are then combined at the final stage to perform calculations through each stream and combine the results of these calculations to obtain a classification decision. The flow at each stage is illustrated in Figure 1.

.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

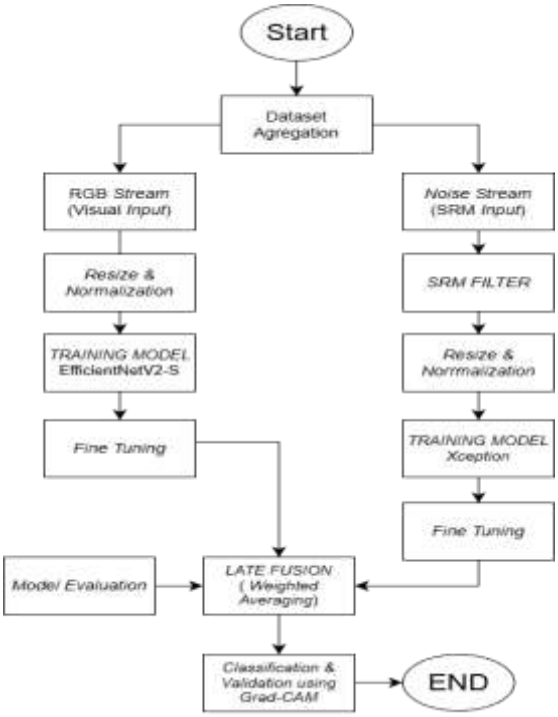

Fig 1 Research Flow

Dataset Agregation

The first stage in the research process is data preparation. The dataset used in this study was obtained from the Kaggle public repository, namely "AI Artwork" (https://www.kaggle.com/datasets/adamelkholy/human-ai-artwork). This dataset contains a collection of human-made paintings and AI-generated paintings taken from various open sources. Kaggle is a platform widely used in academic research because it provides transparent dataset documentation and has been widely used in studies related to generative content detection, so the credibility of the dataset can be accounted for.

Fig 2 Dataset Sample

This dataset consists of 15,000 digital painting images divided into two classes, namely 7,500 original human paintings and 7,500 AI works. The images in the dataset are divided into several painting styles such as impressionism, expressionism, surrealism, realism, baroque, nouveau, ukiyo, and romanticism. During the aggregation stage, all images were collected and grouped into two classes: original paintings and generative AI paintings. The dataset was then divided into three subsets: training data, validation data, and test data, with proportions of 64%, 16%, and 20%, respectively. This division was done to ensure that the training process, model performance monitoring, and evaluation were carried out objectively and did not overlap. Details of the data distribution in each subset are presented in Table 1.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1 Distribusi dataset

No	Dataset	FAKE	REAL	Total
1.	Dataset training	4.800	4.800	9.600
2.	Dataset validation	1.200	1.200	2.400
3.	Dataset testing	1.500	1.500	3.000
	Total	**7.500**	**7.500**	**15.000**

Next, all images are first standardized to a size of 224 × 224 pixels, following the recommended size of the CNN architecture. This size standardization aims to ensure spatial dimension consistency so that the feature extraction process is carried out stably. Then, all images will be converted to RGB color space. This conversion ensures color representation consistency between images. After performing the color conversion, the pixel values are then normalized to the range [0,1] by dividing them by the maximum value of 255. This normalization process aims to stabilize the input value distribution, accelerate the convergence process during training, and prevent the domination of certain scales in the network weight update process. To increase data diversity and reduce the risk of overfitting, data augmentation techniques are applied to the training subset. Augmentation in the RGB Stream focuses on geometric transformations to increase visual display variation without changing the semantic meaning of the painting. In contrast, augmentation in the Noise Stream is limited to simple transformations so that the residual noise patterns are not overly distorted and continue to represent valid generative artifacts. These techniques are presented in Table 2.

Table 2 Augmentation Technique

Input	Configuration
RGB	Rotation, Width Shift, Height Shift, Zoom, Horizontal Flip, Fill Mode
Noise	Horizontal Flip, Random Rotation

The data aggregation stage aims to ensure uniformity in image format and size, as well as compatibility with the EfficientNetV2 and Xception architectures, stabilize the gradient optimization process and accelerate convergence during training through normalization, and increase sample diversity through augmentation techniques, so that the data used is consistent, representative, and unbiased.

RGB Stream Pre-processing

In the Two-Stream Network framework with the Late Fusion strategy, each stream is processed independently so that the pre-processing stages in each stream are designed according to the characteristics of the features to be extracted. In the RGB Stream, all images are first resized to 224 × 224 pixels, following the standard input size in many modern CNN architectures that have been proven stable in extracting multiscale features. After size standardization, pixel values are normalized by mapping pixel intensities from the range [0, 255] to [0, 1] through a rescaling operation. This normalization ensures that the input scale between color channels is within a uniform range, thereby accelerating gradient optimization convergence and preventing numerical instability in the training process. The RGB Stream processing flow then continues with augmentation techniques by applying a data generator that performs direct transformations during training. The augmentation parameter values are limited to a moderate range so that the resulting geometric variations continue to represent natural changes in the angle of view, framing, and visual proportions of the painting, without causing distortions that could significantly alter the composition structure or artistic character. Sequentially, the RGB Stream pre-processing stages include image resizing, normalization, and data augmentation.

This approach aims to enrich the training data distribution so that the model becomes more resistant to geometric variations. As a fair comparison, augmentation is only applied to the training data, while the validation data only undergoes resizing and normalization without geometric transformation, so that the model performance evaluation reflects its generalization ability to the original data distribution.

Noise Stream Pre-processing

Noise Stream is designed to extract visually invisible forensic artifacts by suppressing semantic information and highlighting residual noise patterns generated by neural network-based image generation processes. Several studies have shown that generative AI images leave unnatural correlations between pixels in the high-frequency domain due to convolution, normalization, and upsampling operations within the generator architecture (Zhang et al., 2025). This noise is statistically consistent but difficult to observe directly in the visual domain.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Spatial Rich Models (SRM) are widely used in steganalysis and image forensics due to their ability to extract pixel dependencies and highlight structural components hidden behind the main visual content (Fridrich & Kodovský, 2012). In synthetic image detection studies, several studies have shown that SRM filters are effective in detecting generative AI results because they are able to reveal statistical irregularities in the correlations between pixels generated by the convolution and upsampling processes in visual features (Corvi et al., 2023). Therefore, SRM was chosen as the main approach in Noise Stream to detect characteristic statistical artifacts that distinguish human-made paintings from AI-generated paintings. The initial processing stage was performed by converting the image to grayscale to focus the analysis on microtextures and eliminate color channel redundancy that was irrelevant to noise characteristics. Next, the image is convolved using three 5×5 high-pass SRM kernels, each with a specific function. The KB (3rd-order residue) kernel is used to capture high-frequency texture variations and complex dependencies between pixels, which are sensitive.

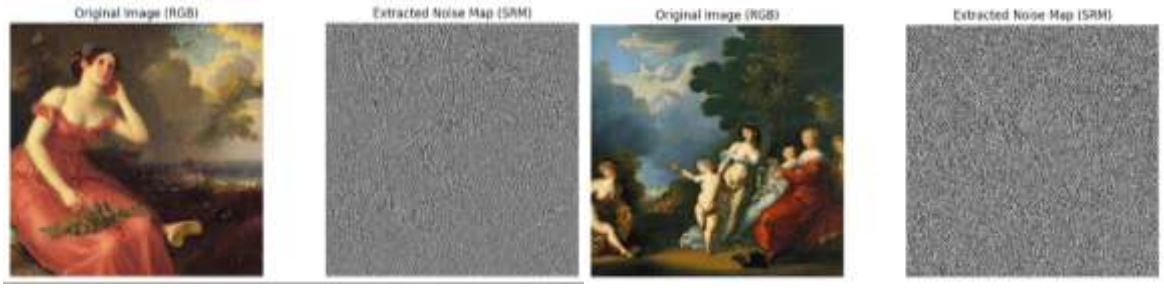

Fig 3 Comparison of visual images and noise

The selection of the KB, KV, and Edge kernel combinations is based on the need to capture the noise variation resulting from the generative AI process across various residual characteristics, ranging from microtexture, local dependence between pixels, to edge discontinuities. This combination allows Noise Stream to obtain a more comprehensive residual representation compared to using a single kernel type, thereby increasing sensitivity to noise patterns typical of generative AI images. To avoid the extreme response of strong edge structures in visual content, a truncation process with a threshold of T = 5 is applied. This technique is commonly used in SRM to suppress residual values and strengthen weak noise signals that are more relevant as forensic traces (Fridrich & Kodovský, 2012). The residual map is then normalized to the range [0, 255] to be compatible as CNN network input. In addition to residual extraction, augmentation in the form of random JPEG compression, scaling, small rotations, and flipping was applied. This technique allows the model to be trained to recognize noise patterns that remain consistent even when image quality and geometric transformations vary (Corvi et al., 2023). The main objective of Noise Stream is to extract noise artifacts left behind by AI-based image generation processes, by suppressing visual semantic information and highlighting noise patterns as a supporting factor in the detection of human-made paintings and generative AI paintings. In summary, the Noise Stream workflow begins with converting images to grayscale to remove color information, followed by noise extraction using an SRM filter. The resulting noise is then controlled through a truncation process to suppress extreme values, normalized to accelerate convergence, and amplified through augmentation. Within the Two-Stream Network framework with a Late Fusion strategy, the noise representation is then combined with visual features from the RGB Stream so that the system can improve accuracy, robustness, and reliability in detecting generative AI paintings.

RGB Stream Architecture

In the RGB Stream path, the EfficientNetV2 architecture was selected as the main backbone. EfficientNetV2 was chosen based on its ability to balance visual representation performance and computational efficiency. EfficientNetV2 has a better accuracy-to-parameter ratio than conventional CNNs, making it suitable for training scenarios with limited resources. After the image data was processed in the pre-processing stage. Next, the entire EfficientNetV2 backbone layers are frozen, and only the classification layer is trained. The features extracted by the backbone are summarized using Global Average Pooling, followed by Batch Normalization to stabilize the activation distribution. A fully connected layer of 512 neurons with ReLU activation is used to increase discriminatory capacity, while high-ratio dropout is applied as the main regularization mechanism so that no neuron becomes too dominant in learning.

In the second stage, all backbone layers are reopened and fine-tuning is performed using a very small learning rate. This approach allows the model to adjust visual feature representations to the specific characteristics of AI-generated paintings, such as visual distortion and structural inconsistencies, without damaging the basic features learned during the pre-training stage. The training process is combined with cosine learning rate

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

scheduling to gradually adjust the learning rate decrease, as well as validation loss-based early stopping to control overfitting and ensure stable convergence.

Noise Stream Architecture

Noise Stream is designed to analyze noise characteristics that are not visually apparent but are forensically relevant. The input to this stream is a noise map extracted using the Spatial Rich Models (SRM) approach. SRM is applied as a set of fixed high-pass filters that aim to suppress semantic information and highlight high-frequency noise patterns associated with generative processes and post-processing stages of images. The noise extraction process begins with converting the image to grayscale to focus the analysis on noise and texture. Next, the image is convolved using several 5×5 SRM kernels designed to capture residual texture variations, subtle noise disturbances, and structural artifacts around object edges. The convolution results from each kernel are combined into a three-channel image to be compatible with the CNN model input. To improve learning stability, a truncation process is applied to limit extreme values, followed by normalization to keep the noise distribution within a controlled range. The Xception backbone is used in Noise Stream because of its ability to model spatial correlations through a depthwise separable convolution mechanism. Xception, with depthwise separable convolution, reduces the number of computational operations without sacrificing noise feature extraction capabilities. Compared to conventional CNN architectures, Xception is able to extract fine textural and statistical features more efficiently, making it a relevant choice for noise domains.

The Noise Stream training process is also carried out in two stages. In the initial stage, all Xception backbone layers are frozen and training is focused on the classification layer to stabilize initial convergence. In the next stage, all layers are reopened and fine-tuned using a low learning rate so that the model can adapt to the specific noise characteristics in AI-generated paintings without increasing the risk of overfitting. In addition, Noise Stream is equipped with image degradation-based augmentation, such as JPEG compression, resolution changes, and light geometric transformations. This approach aims to improve the model's resilience to common manipulations and prevent overfitting to specific noise patterns. Although Noise Stream's individual performance is lower than RGB Stream, this path serves as a complement in the late fusion stage, providing additional information beyond the visual domain, thereby strengthening the system's ability to detect AI images with high visual quality but statistical inconsistencies.

Late Fusion

In the Late Fusion combination stage, the outputs from the RGB Stream and Noise Stream are combined at the decision level to produce the final classification probability. Each model generates probability values through the Sigmoid activation function in the output layer in a probability range between 0 and 1. The combination is performed using the weighted averaging method commonly used in classifier systems. It is formulated as follows:

$$P_{final} = \alpha P_{rgb} + (1 - \alpha) P_{noise}$$

The weighted average formulation is used with the assumption that the sigmoid output in each stream represents a probability estimate. In the context of decision-level fusion, linear combination of probabilities is a commonly used approach because it is stable, easy to interpret, and effective when each classifier works on different feature representations and does not share learning parameters directly (Kuncheva, 2014). In the equation above, P_rgb and P_noise represent the output probabilities of the RGB Stream and Noise Stream, respectively. The parameter α serves as a weight coefficient that regulates the relative contribution of each stream to the final decision. The selection of the weighted averaging approach was based on design and model stability considerations. The RGB Stream tends to have stronger discriminatory capabilities in terms of semantic visuals, while the Noise Stream acts as a complementary path that captures residual statistical inconsistencies. Therefore, to ensure that the final decision does not depend on a specific weight configuration, a sensitivity analysis of the α parameter was performed by testing several weight combinations on the validation data. The test was conducted on α values ranging from 0.5 to 0.8 at specific intervals to observe changes in classification performance.

As an illustration of the probability combination mechanism, suppose the weight is set to α = 0.6, which indicates that the RGB Stream contributes 60% to the final decision, while the Noise Stream contributes 40%. Given the output probability values Prgb = 0.85 and Pnoise = 0.40, the final probability is calculated as follows:

$$P_{final} = (0.6 \times 0.85) + (0.4 \times 0.40)$$
$$P_{final} = 0.51 + 0.16 = 0.67$$

Using a decision threshold of 0.5, the final probability value of 0.67 results in a decision that the image is classified as the Real class, because the value is above the threshold. This example shows how the weighted averaging method allows both streams to contribute explicitly to the decision-making process. Although this

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

approach assumes independence between the RGB Stream and Noise Stream outputs, the assumption of independence between the RGB Stream and Noise Stream outputs is used as a practical approach given that the two streams process fundamentally different representations, namely visual features and noise.

Evaluation and Model Intepretability

The final stage of the research was conducted through a comprehensive evaluation of model performance and interpretability analysis, which included quantitative evaluation, decision sensitivity analysis, and qualitative evaluation based on Explainable AI. Quantitative evaluation was performed using accuracy, precision, recall, and F1-score metrics, accompanied by confusion matrix analysis to observe the distribution of classification errors between original painting images and AI-generated images. Accuracy, precision, recall, and F1-score metrics were chosen because they are quantitative indicators. Accuracy is used to measure the overall performance of the model, precision represents the confidence level of predictions for positive classes, recall measures the model's ability to detect all positive samples, while F1-score is used as an aggregate metric to balance precision and recall.

The model used is a Two-Stream approach with a Late Fusion mechanism at the prediction score level, where the probability outputs from the RGB Stream and Noise Stream models are combined using specific weightings before a final decision is made. To ensure that the model's performance does not depend on a specific configuration, a sensitivity analysis of the fusion weight was performed. A quantitative evaluation was also performed using Receiver Operating Characteristic (ROC) analysis and Under the Curve (AUC) values were calculated separately for each independent model, namely the RGB Stream and Noise Stream. ROC analysis was used to assess the model's discriminatory ability in distinguishing between the two classes at various decision thresholds, without affecting the main evaluation results on the Two-Stream model. The high AUC value on the RGB Stream indicates that the EfficientNetV2 backbone has excellent class separation capabilities in the visual domain, while the ROC on the Noise Stream provides an overview of the contribution of the noise-based forensic pathway. Qualitative evaluation was performed using an Explainable AI approach through the Gradient-weighted Class Activation Mapping (Grad-CAM) method. This technique visualizes the image areas that are the focus of attention of the model when generating classification decisions in the form of a heatmap. Visualization is performed on a number of test data samples to observe the image areas that are the focus of attention of the model in generating classification decisions. This visualization is used as a qualitative interpretation tool to verify in general that the model tends to utilize visual content and noise. However, this analysis does not yet include a systematic study of false prediction cases, so mitigation of the Clever Hans Effect phenomenon is still indicative and constitutes a limitation of this study.

RESULT

Configuration and Model Training

The model training was designed to evaluate the effectiveness of the Two-Stream architecture with the Late Fusion mechanism through the same training configuration on both Stream paths. Before the final configuration was determined, several training scenarios were tested with variations in the number of epochs, layer freezing schemes, and learning rate reduction strategies to obtain a configuration that provided stable convergence and the best validation performance. The configuration reported in this study is the configuration with the most stable performance based on the tests that have been conducted. Training was carried out using a two-stage transfer learning approach as commonly used in EfficientNet-based image classification tasks (Tan & Le, 2021). In the first stage, the model backbone was frozen and training was conducted for 15 epochs by only updating the classification layer so that the general feature representation of the pre-trained model was maintained while adjusting the classifier to the target data. The next stage involved a fine-tuning process for 20 epochs by opening all layers of the network so that the model could adjust the feature representation to the characteristics of the paintings and AI-generated images.

The number of fine-tuning epochs follows common practice in transfer learning research, which shows that additional training of around 15–30 epochs is sufficient to improve performance without significantly increasing the risk of overfitting (He et al., 2016). The learning rate setting strategy uses a cosine annealing mechanism that gradually decreases the learning rate from 0.001 to 0.0001 to maintain optimization stability, then continues with a small learning rate of 1 × 10-5 at the fine-tuning stage to avoid extreme weight changes that can cause catastrophic forgetting (Loshchilov & Hutter, 2017). Training uses the Adam optimizer with a batch size of 32, which is commonly used in modern CNN training because it provides a balance between gradient stability and computational efficiency. This configuration is applied identically to the RGB Stream and Noise Stream before being combined in the Late Fusion stage. The complete details of the hyperparameters used are presented in Table 2.

Table 3 Hyperparameter Pengujian

Parameter	Settings

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Batch Size	32
Optimizer	Adam
Input Image Size	224 x 224
Dropout Rate	0.6
Dense Layer	512 (RGB Stream), 256 (Noise Stream)
Training Epochs	35
Learning Rate	1×10^{-5}
Loss Function	Binary Crossentropy
Scheduler	Cosine Annealing

The learning rate value of 1×10^{-5} in Table 2 represents the learning rate in the fine-tuning phase, while the initial training phase uses a larger learning rate that is gradually reduced through cosine annealing. Model training is carried out separately for each stream before being combined at the final stage. Each stream underwent an analysis of the architecture used so that the training configuration would match the characteristics of the visual feature representation and residual noise being learned.

Single Stream Model Performance Evaluation

Initial testing was conducted on each stream separately as a baseline. On the RGB stream, four variants of EfficientNetV2 (B0, B1, B2, S, M) were compared to determine the best backbone for extracting painting features. Evaluation was performed using accuracy and loss metrics on the test and validation data, as summarized in Table 3.

Table 4 Performance Comparison of EfficientNetV2 Variants

Arsitektur	Accuracy	Loss	Valid Accuracy	Valid Loss
EfficientNetV2-B0	0.98	0.052	0.98	0.045
EfficientNetV2-B1	0.98	0.048	0.98	0.033
EfficientNetV2-B2	0.98	0.05	0.99	0.023
EfficientNetV2-S	**0.99**	**0.014**	**0.98**	**0.034**
EfficientNetV2-M	0.99	0.026	0.98	0.071

The results show that increasing model complexity contributes to learning stability. EfficientNetV2-S produced the lowest loss and highest accuracy, so it was selected as the backbone for the visual pathway. The training accuracy and loss curves are shown in Figure 4.

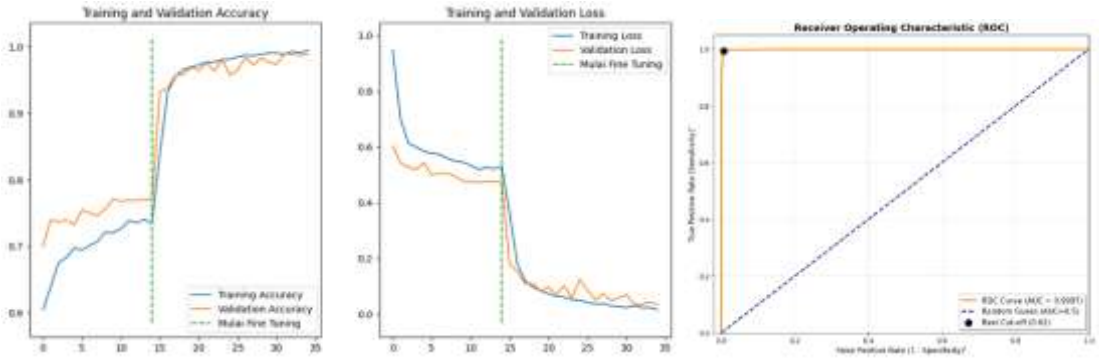

Fig 4 Accuraty dan Loss Model EfficientNetV2-S

As an additional analysis of the model's discriminatory ability, ROC and AUC evaluations were performed to assess the model's discriminatory ability independently of the decision threshold. On the RGB Stream with the EfficientNetV2-S backbone, an AUC value of 0.9997 was obtained. The optimal decision threshold was at 0.6223, which provided the best balance between sensitivity and specificity. As a comparison, the Noise Stream path was trained using the Noise Map resulting from Spatial Rich Models (SRM). The Xception model produced an accuracy of 0.90 with a stable loss reduction of 0.24 in the fine-tuning phase.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

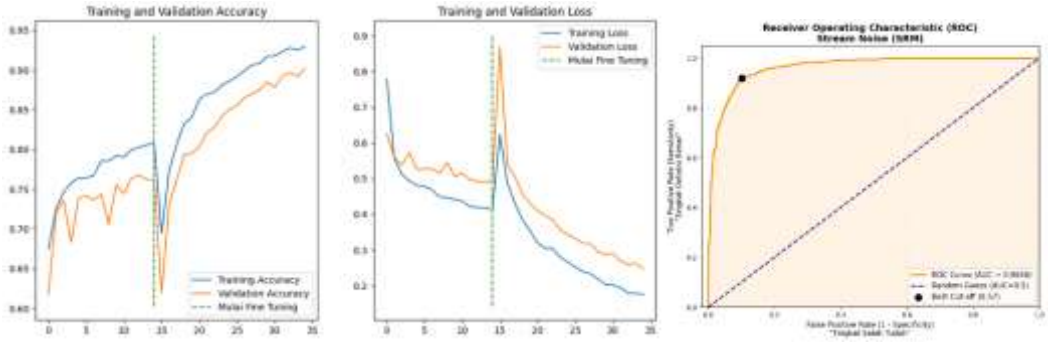

Fig 5 Accuracy dan Loss Model Xception

Meanwhile, on the Xception-based Noise Stream, it obtained an AUC value of 0.9646, indicating that the model was able to detect noise patterns. The optimal threshold value of 0.5665 showed that the noise path had a significant discriminatory contribution even though its performance was lower than that of the visual path. The selection of Xception was based on its ability to extract low-level statistical patterns and textural features. (Akbar, 2025).

The difference in performance between the two paths indicates that visual features and noise features have complementary discriminatory characteristics. The RGB stream excels at capturing semantic and morphological information of objects, while the noise stream is effective at detecting statistical inconsistencies that are often overlooked by AI generative processes. To ensure that the performance difference between the RGB and Noise Streams is not random, statistical significance testing was performed using the McNemar test. This test was chosen because both models were evaluated on the same test data set, so the prediction results were paired. The McNemar test is a non-parametric test used to compare two classification models with a focus on differences in prediction decisions for each sample. The test results produced a statistical value of 180.57 with a p-value of 3.65×10^{-41}, which is much smaller than the significance threshold of 0.05. This very small p-value indicates that the difference in performance between the two paths is statistically significant. This finding indicates that the two paths extract different feature characteristics and are complementary to each other.

Two Stream Model Performance Evaluation

In the Two-Stream Late Fusion evaluation stage, the best models from each stream, namely EfficientNetV2-S on the RGB Stream and Xception on the Noise Stream, are combined at the prediction score level using a weighted averaging strategy. This approach combines the output probabilities of both models before a final decision is made, allowing for flexible integration of visual and residual noise information. In its implementation, two main weighting configurations were evaluated. The configuration with balanced weighting between the RGB and Noise Streams (0.5 : 0.5) is referred to as Model B, while the configuration with weighting dominated by the RGB Stream (0.7 : 0.3) is referred to as Model A. A summary of the performance of each configuration is presented in Table 4.

Table 5 Model Comparison Performance

Arsitektur	Accuracy	Precision		Recall		F1-Score	
		Real	Fake	Real	Fake	Real	Fake
Xception	0.90	0.93	0.88	0.87	0.94	0.90	0.91
EfficientNetV2-S	0.99	1.00	0.99	0.99	1.00	1.00	1.00
Xception + EfficientNetV2 *Two Stream* A	**0.98**	**1.00**	**0.98**	**0.98**	**1.00**	**0.99**	**0.99**
Xception + EfficientNetV2 *Two Stream Late* B	0.92	0.99	0.87	0.85	0.99	0.92	0.93

The results in Table 4 show that the Two-Stream Late Fusion approach can improve model accuracy compared to a single Noise Stream, particularly in terms of the Fake class recall metric, which is a crucial aspect in the context of detecting AI-generated images. However, balanced fusion (0.5:0.5) produces lower performance than the single RGB model. This is due to the contribution of the Noise Stream, which has lower performance, so

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

that when the weights are balanced, the less stable predictions from the noise stream can reduce the final decision that is actually correct in the visual stream. Conversely, the configuration with RGB weight dominance shows better performance. To evaluate the influence of weights and decision thresholds more systematically, a sensitivity analysis was conducted on variations in fusion weights and classification thresholds. The results of testing several configurations are presented in Table 5.

Table 6 Sensitivity Analysis of Two-Stream Late Fusion Weights and Thresholds

Weight RGB	Threshold	Accuracy	F1-Score
0.5	0.5	0.92	0.93
0.6	0.6	0.99	0.99
0.7	**0.7**	**0.997**	**0.997**
0.7	0.5	0.99	0.99

The sensitivity analysis results show that Two-Stream performance is relatively stable in the medium to high weight range. The configuration with an RGB weight of 0.7 and a threshold of 0.7 produced the highest performance with an accuracy of 99.7% and an F1-Score of 0.997. Several other configurations, such as weights of 0.6 and 0.7 with corresponding thresholds, also showed very similar performance. These findings indicate that the improvement in Two-Stream performance does not depend on a specific combination of parameters, but rather reflects the robust integration of visual features and noise against variations in fusion parameters.

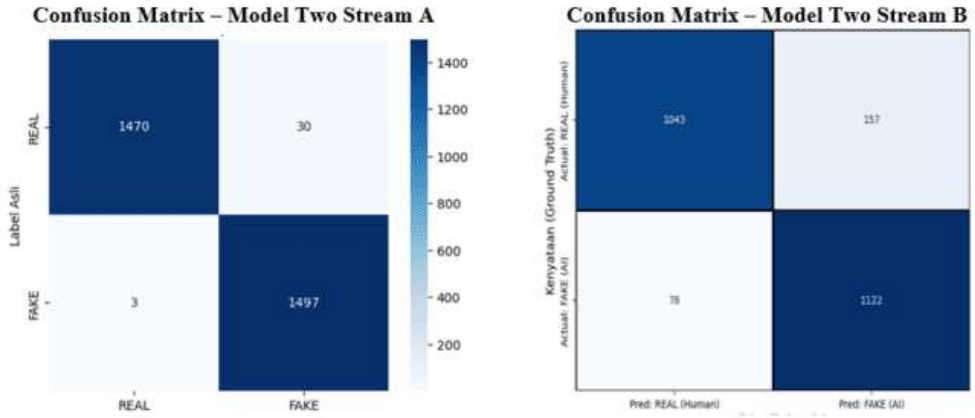

Fig 6 Confusion Matrix Model *Two Stream* (Model A, B)

The distribution of classification errors was further analyzed using a confusion matrix to clarify the implications of the fusion configuration on the types of classification errors. The visualization shows that the configuration with RGB weight dominance consistently reduces the number of false negatives, especially in the Fake class. To ensure that the differences in performance between the Two-Stream configurations were not solely due to random variations, statistical significance testing was performed using the McNemar Test. This test was applied by comparing the configuration with the highest performance, namely RGB weight 0.7 and threshold 0.7, against an alternative configuration with lower performance, namely RGB 0.5 and threshold 0.5. The test results showed a value of n10 = 8, which represents the number of samples that were correctly classified by the best configuration but incorrectly classified by the comparison configuration, and n01 = 9 for the opposite condition. The p-value obtained was 1.0. This p-value indicates that the observed improvement in numerical performance was not accompanied by a statistically significant difference in classification decisions at the sample level. Thus, the increase in accuracy from 0.92 to 0.98 in the Two-Stream configuration cannot be claimed as a statistically significant improvement, but rather as an empirical improvement that shows a tendency for performance improvement. This evaluation shows that the Two-Stream Late Fusion approach provides improved reliability and stability of detection, mainly through the reduction of false negatives and consistency of performance against variations in fusion parameters.

Model Interpretability

In addition to performance evaluation based on Confusion Matrix and Classification Report, this study also conducted model interpretability analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) to validate the transparency and rationality of classification decisions. This approach aims to identify image areas that have the most significant contribution to model predictions. In the Two-Stream Late Fusion architecture,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

visualization is performed separately on the RGB Stream and Noise Stream to display the focus characteristics of each path.

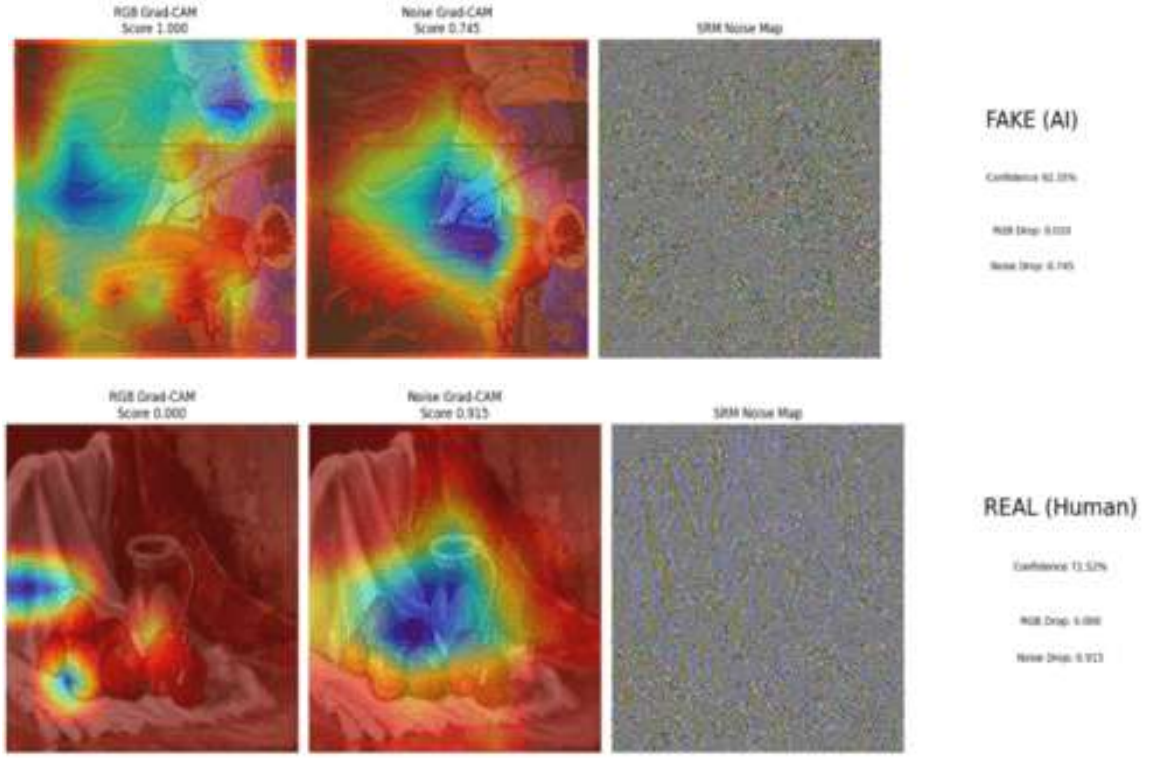

Fig 7 Visualization of Grad-CAM Results

Figure 7 shows examples of Grad-CAM visualization on FAKE and REAL images, along with noise maps generated by the Spatial Rich Model (SRM). In the RGB Stream, the Grad-CAM heatmap shows a focus on semantic visual elements such as painting textures, main object areas, and color transitions. The activation tends to follow the visual structure of the image, both in the FAKE and REAL classes. Conversely, in the Noise Stream, the activation pattern does not follow the shape of the object, but is scattered in certain areas related to the distribution and density of residual noise highlighted by the SRM.

To complement the visual analysis, faithfulness was evaluated using an occlusion-based confidence drop approach by removing the image areas with the highest Grad-CAM values in each stream. In the FAKE image generated by AI, the model initially produced a classification rate of 92.35%. After the Grad-CAM areas were removed, the confidence drop in the RGB Stream was relatively small (RGB Drop = 0.010), while in the Noise Stream, there was a much larger drop (Noise Drop = 0.745). These results indicate that the model's decision on AI images is more sensitive to features extracted in the noise stream than to semantic visual features. Conversely, in REAL images with a classification rate of 72.52%, removing the Grad-CAM area in the RGB Stream did not cause any change in confidence (RGB Drop = 0.000), while removing the area in the Noise Stream resulted in a very large decrease in confidence (0.915). This finding indicates that even when images are classified as REAL, the noise path still plays an important role in maintaining the model's decision. Overall, these interpretability results show that the RGB Stream functions in capturing visual and semantic information, while the Noise Stream provides a more dominant and consistent contribution in distinguishing AI-generated images and human images through residual noise characteristics. However, the evaluation of interpretability in this study is still limited to one explanation method, namely Grad-CAM, and one occlusion-based faithfulness evaluation approach.

Computanional Cost Analysis

A computational cost analysis was conducted to evaluate the implications of using the Two-Stream Late Fusion approach compared to the Single Stream RGB and Noise models. This evaluation was based on empirical evidence in the form of the number of parameters, FLOPs, inference time, and GPU memory usage obtained from direct model profiling.

Table 7 Comparison Computational Cost Model

Metric	RGB	Noise	Two-Stream
Params (M)	20.99	21.39	42.39
FLOPs	5.75	9.14	14.89

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15814
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Inference Time (ms)	79.31	71.51	150.82
GPU Memory (MB)	3447.90	3447.90	6895.80

Based on the results in Table 6, the RGB and Noise models have a relatively comparable number of parameters, each in the range of 21 million parameters. This shows that architecturally, the two models have almost equivalent complexity. However, differences begin to appear in computational requirements, where Noise Stream requires higher FLOPs compared to RGB Stream. This finding is consistent with the characteristics of noise-based feature extraction, such as Spatial Rich Models, which emphasize statistical analysis and low-level textures, thus requiring more intensive convolution operations.

Figure 4. Computational Cost Comparison

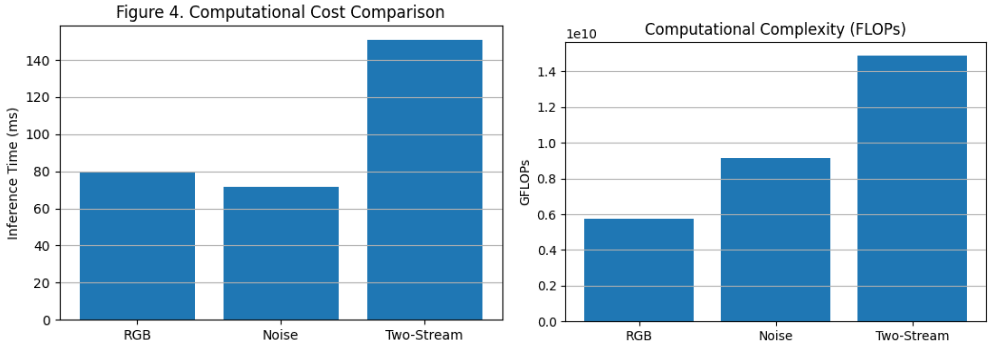

Fig 8 Computing Cost Comparison Chart

Based on Figure 8, the Two-Stream approach shows a significant increase in all aspects of computational cost. The number of parameters is almost doubled compared to the single model because both backbones are run in parallel. The same is reflected in the FLOPs value, which is a direct accumulation of RGB and Noise Stream. The visualization in the comparison makes it clear that Two-Stream requires longer inference time. In terms of GPU usage, the Two-Stream model requires significantly more memory than the single model. This is due to the simultaneous reinforcement of two models during the inference process.

DISCUSSIONS

The results of this study show that the Two-Stream with Late Fusion approach, which integrates visual information and residual noise, is effective in distinguishing between human-made and AI-generated paintings. Experiments prove that the weighted fusion strategy produces more stable performance than balanced fusion. The configuration with weight dominance in the visual stream is able to achieve high accuracy that is close to the best Single-Stream model. Sensitivity analysis on weight variations and decision thresholds shows that Two-Stream performance does not depend on a single configuration. Several combinations of weights and thresholds produce consistently high performance, indicating that the model not only works optimally under specific conditions but also has prediction stability against changes in Late Fusion parameters. To ensure that the performance differences between Two-Stream configurations did not occur by chance, statistical significance testing was performed using the McNemar test on the main configuration pairs. The test results show that the differences between the best configuration and several comparison configurations are statistically significant, while the differences between configurations with similar performance are not significant. The advantage of Noise Stream lies in its ability to detect statistical artifacts left behind by generative AI processes, regardless of artistic style. Noise resulting from convolution and upsampling operations is not always visually apparent, but is exposed through Spatial Rich Models filters. Grad-CAM visualization shows that each stream learns different representations. RGB Stream focuses on morphological and visual structure irregularities, while Noise Stream highlights areas with high noise density that do not follow the shape of objects. In terms of efficiency, the Two-Stream approach has the consequence of increased computational costs. However, this increase in cost is in line with the improvement in model performance, reflecting a reasonable trade-off between performance and efficiency. Thus, this approach is more suitable for forensic or content validation scenarios, where accuracy and reliability are prioritized over low latency. Although the results are promising, this study still has limitations. Generalization evaluation was conducted indirectly through noise and interpretability analysis, so cross-dataset or cross-domain testing is still needed to confirm the model's overall robustness.

CONCLUSION

Based on a series of experiments, this study concludes that the Two-Stream Late Fusion Network architecture is an effective and reliable approach for detecting paintings produced by generative AI. The separation of feature extraction into two independent streams, namely the RGB Stream based on EfficientNetV2-S and the Noise Stream based on Xception with Spatial Rich Models (SRM) input, has been proven capable of capturing

*name of corresponding author

complementary characteristics of AI images. The visual stream excels in semantic information, while the noise stream is effective in detecting statistical areas and noise that are not always visible. The evaluation results show that the Single-Stream EfficientNetV2-S model achieved the highest performance with 99% accuracy and an F1-Score of 1.00, while Xception had lower performance with 90% accuracy and an F1-Score of 0.91. In the Two-Stream configuration, the Late Fusion strategy with appropriate weighting was able to achieve good performance while improving detection stability. The best configuration with RGB weight 0.7 and threshold 0.7 resulted in 99.7% accuracy and an F1-Score of 0.997, while balanced weighting lowered performance to 92% accuracy, demonstrating the importance of visual feature dominance with noise support as a complementary mechanism. In terms of computational efficiency, the Two-Stream model has 42.39 million parameters, almost double that of using RGB Stream or Noise Stream alone. Computational complexity also increases in terms of FLOPs, inference time, and GPU memory in the model. These findings confirm a clear trade-off between increased accuracy and computational cost. Although the results show high and stable performance, this study still has limitations. The evaluation was conducted on a single dataset, so the model's cross-dataset generalization ability has not been empirically validated. Furthermore, the computational efficiency analysis is still descriptive and has not explored architectural optimization comprehensively. Therefore, further research is recommended to conduct cross-dataset evaluation to test the model's generalization robustness, develop adaptive or dynamic fusion strategies, and perform computational efficiency analysis through the use of lighter architectures or model optimization techniques. Overall, the proposed Two-Stream Late Fusion method has strong potential as the foundation for AI-based digital art authentication systems, with a balance between accuracy and computational efficiency. Overall, the proposed Two-Stream Late Fusion method has strong potential as the foundation for AI-based digital art authentication systems, with a balance between high accuracy, robustness, and interpretability.

REFERENCES

Adventino Gulo, S., Amelia Pertiwi, A., Putri Syaifullah Nasution, S., & Syahputra, H. (2025). Deteksi Deepfake Dalam Citra Menggunakan Convolutional Neural Network (Cnn). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(5), 8655–8660. https://doi.org/10.36040/jati.v9i5.14896

Akbar, M. H. (2025). *Forensik Citra Digital Berbasis XceptionNet dengan Kerangka Kerja DFRWS untuk Deteksi Deepfake*. *xx*(xx), 221–228.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Anggraini, D., Handayaningrum, W., Rahayu, E. W., Suryandoko, W., & Sabri, I. (2024). Kolaborasi seniman dan kecerdasan buatan (AI) dalam membangkitkan gelombang kreativitas di era revolusi seni digital. *Imaji: Jurnal Seni Dan Pendidikan Seni*, *22*(2), 111–119. https://doi.org/10.21831/imaji.v22i2.69734

Aris, S., Aeini, B., & Nosrati, S. (2023). A Digital Aesthetics? Artificial Intelligence and the Future of the Art. *Journal of Cyberspace Studies*, *7*(2), 219–236. https://doi.org/10.22059/JCSS.2023.366256.1097

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. In *Multimedia Systems* (Vol. 16, Issue 6). https://doi.org/10.1007/s00530-010-0182-0

Bianco, T., Castellano, G., Scaringi, R., & Vessio, G. (2023). *Identifying AI-Generated Art with Deep Learning*. *October*.

Castellano, G., Grazia Miccoli, M., Scaringi, R., Vessio, G., & Zaza, G. (2024). Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps. *CEUR Workshop Proceedings*, *3839*, 65–74.

Cetinic, E., Lipic, T., & Grgic, S. (2018). Fine-tuning Convolutional Neural Networks for fine art classification. *Expert Systems with Applications*, *114*, 107–118. https://doi.org/10.1016/j.eswa.2018.07.026

Fridrich, J., & Kodovský, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *7*(3), 868–882. https://doi.org/10.1109/TIFS.2012.2190402 He, K. (2015). *Deep Residual Learning for Image Recognition*.

Kuncheva, L. I. (2014). *Combining pattern classifiers: Methods and algorithms* (2nd ed.). Wiley.

Li, M., & Stamp, M. (2025). *Detecting AI-generated Artwork*. Detecting AI-generated Artwork. *arXiv preprint arXiv:2504.07078*. https://arxiv.org/abs/2504.07078

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. *Proceedings of the International Conference on Learning Representations (ICLR)*. Mahara, A., & Rishe, N. (2025). *Methods and Trends in Detecting AI-Generated Images : A Comprehensive Review ⋆*. *3552*, 1–35.

Morariu, V. I., & Davis, L. S. (n.d.). *Two-Stream Neural Networks for Tampered Face Detection*.

Nasir, A., & Tariq, Z. A. (2024). Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification. *IEEE Access*, *12*(October), 184119–184131. https://doi.org/10.1109/ACCESS.2024.3503413

Simonyan, K. (n.d.). *Two-Stream Convolutional Networks for Action Recognition in Videos arXiv : 1406 . 2199v2*

*name of corresponding author

[cs . CV] 12 Nov 2014. 1–11.

Subkhi, M. B., Setiawan, A. B., & Candra, M. Y. A. (2023). Klasifikasi Gambar: Membedakan Lukisan Buatan Manusia dan AI dengan CNN. *Paradigma: Jurnal Filsafat, Sains, Teknologi, Dan Sosial Budaya*, *29*(4), 149–155.

Tan, M., & Le, Q. V. (2021). *EfficientNetV2 : Smaller Models and Faster Training*.

Vivekananda, G. N., Mahesh, T. R., Gupta, M., Thakur, A., & Sayal, A. (2025). Refining digital security with EfficientNetV2-B2 deepfake detection techniques. *Egyptian Informatics Journal*, *30*(February), 100699. https://doi.org/10.1016/j.eij.2025.100699

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). *How transferable are features in deep neural networks ? 27*.

Zhang, Y., Pang, Z., Huang, S., Wang, C., & Zhou, X. (2025). Unmasking AI-created visual content: a review of generated images and deepfake detection technologies. *Journal of King Saud University - Computer and Information Sciences*, *37*(6). https://doi.org/10.1007/s44443-025-00154-8

*name of corresponding author