IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 3, January 2026

Impact Factor: 7.67

Explainable AI (XAI) for Forensic Analysis: Image Forgery Detection and Deepfake Video Identification

Chandraiah T[1] and Sumanashree Y S[2]

Assistant Professor, Department of Computer Science, Yuvaraja College, Mysore[1]
H.O.D and Assistant Professor, Post Graduate Department of Computer Science
J. S. S. College of Arts, Commerce and Science, Mysore[2]
chandruycm@gmail.com and sumanashreeys@gmail.com

Abstract: *Their most recent development has made it possible to create high-quality forgeries and deepfakes videos: the swift advancement of generation models, including Generative Adversarial Networks (GANs) and diffusion models, has facilitated their production. The digital forensics, legal evidence validation and trust of the people is under serious challenge because they can not be easily detected by the customary forensic methods. Detecting such manipulations with deep learning methods has high accuracy due to convolutional and recurrent neural networks. Nevertheless, their opaque, black-box character makes them less applicable in forensic as well as judicial settings, where transparency, interpretability and traceable evidence are vital.*

The study presents an Explainable Artificial Intelligence (XAI)-driven system of forensic analysis, which is a combination of CNN-based image forgery detection and hybrid CNN-LSTM-based deepfake video detection. The framework uses the post-hoc and intrinsic XAI methods including Grad-CAM, LIME, and SHAP to produce human-readable explanations on both visual and feature scales. Through experimental assessment on benchmark datasets, it is shown that the proposed XAI-enhanced models not only attain competitive levels in detection but also present understandable evidence. Such perfect accuracy and transparency make the framework applicable to forensic inquiry and legal proceedings..

Keywords*:* Explainable AI, Digital Forensics, Image Forgery Detection, Deepfake Videos, CNN, XAI

I. INTRODUCTION

The digital media has become a powerful medium of communication, information sharing, entertainment as well as legal documents. The popularity of smart phones, social media websites, and the creation of digital content has greatly augmented the quantity and availability of digital images and videos. Although this proliferation has made the content creation more democratic, the content has also created serious problems of authenticity and trust. The digital media has become highly sophisticated, especially in the domain of image forgery and deepfakes videos, and it is now quite complex to ensure that human observers or the use of conventional forensic practices can consistently identify the changes.[1] These manipulations are not the technical issues but the ones that have a severe impact on the trust in society and politics, as well as legal evidence and cybersecurity. These vulnerabilities can be exploited by fake news, impersonation attacks, and fraudulent media during judicial proceedings and thus a strong detection mechanism is urgently required.[2]

The development of image editing software and generative models, in particular, Generative Adversarial Networks (GANs) and diffusion models, have made it possible to create highly realistic forged images and deepfake videos.

[1] Farid, H. (2019). *Digital image forensics: There's more to images than meets the eye*. MIT Press.

[2] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53.

DOI: 10.48175/IJARSCT-30926

GAN-based models are trained to produce synthetic images in opposition to the discriminator network, whereby the generator network produces content that can be close to real-world characteristics (facial expression, lighting, textures, and background consistency, etc.). Likewise, the diffusion models repeatedly optimize noisy input data to generate high-quality output that is capable of the convincing reproduction of natural images and videos. Such generative methods are now faster than other traditional forensic methods, which typically use features created by hand, like noise pattern inconsistencies, JPEG compression artifacts, lighting patterns, or pixel anomalies. Consequently, the modern digital forensics is confronted with the two-fold challenge of identifying any manipulations and, at the same time, ensuring reliability and interpretability of the evidence that can be further used in the investigative or judicial process.[3]

Detection techniques based on deep learning have demonstrated impressive capability in handling such problems. Convolutional Neural Networks (CNNs) have demonstrated themselves to be useful in detecting spatial inconsistencies in images, such as subtle texture variations, edge distortions and anomalies in color distribution that are indicative of manipulation. In the case of videos, frame-to-frame errors, unnatural blinking, or non-existing motion patterns can be trained with hybrid CNNLong Short-Term Memory (LSTM) systems, and deepfakes that use facial and audio synchronization can be detected. Such models tend to be very accurate, performing better than other forensic approaches, and can be extended to forms of manipulation which have never been seen before.

Regardless of their performance, deep learning models are often criticized to be black boxes, where one has a slight idea as to why a specific decision was arrived at. This obscurity is a major detriment in forensic settings, since evidence given in court must be understandable, readable and defensible. Such unaccountability may result in doubts about the trustworthiness of automated detection systems and this may be an impediment to the use of automated detection in legal practice.[4] An example of this is that a CNN can raise a red flag on an image as fake, without providing interpretative arguments such as heatmaps of areas of manipulation or feature importance scores, it is hard to trust the model to be right. On the same note, deepfake video classification should be clear on what frames or time disparities were used in temporal analysis to be classified.

Explainable Artificial Intelligence (XAI) is a solution to this weakness by introducing transparency and interpretability to AI models. XAI methods can be divided into intrinsic algorithms which integrate interpretability into the model architecture and post-hoc algorithms which analyze the trained models to give an explanation of their prediction. Typical methods of XAI in forensic analysis are Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), and Shapley Additive Explanations (SHAP). Grad-CAM identifies the areas of the image or video frame that the model relied on most, LIME offers instance-level explanations, by seeking to approximate the decision boundary in the area, and SHAP quantifies the contribution of individual features to the prediction. The use of such techniques allows visual and quantitative imaging of why a picture or video is labeled as being manipulated by forensic investigators, which adds credibility to automated detection tools and adds greater evidence value to them.

The application of XAI to forensic detection systems is especially applicable to areas of the law and investigation. The growing demands of courts and regulatory bodies are that AI-related tools need to be explainable, accountable, and ethically implemented. Explainable forensic models, in addition to supporting trust and adoption, can also make AI systems and human experts to work more effectively. Algorithms can be used to complement domain knowledge in the work of the investigator to enhance the precision and accuracy of digital evidence analysis.[5] Furthermore, interpretability facilitates cross-validation of model choices among datasets and manipulation modalities, a requirement due to the fast changing generative methodologies that generate novel forgery patterns that never existed before.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

[4] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1–11).

[5] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

DOI: 10.48175/IJARSCT-30926

In this paper, we will discuss how XAI can be applied to forensic image forgery detection and deepfake video recognition. It presents a unified architecture integrating high-precision deep learning with the most recent XAI methods to compromise between the performance of detection and the ability to interpret. The paper justifies the suggested framework by testing it on benchmark data and showing that XAI-improved models can attain competitive detection accuracy and give interpretations acceptable to humans both at visual and feature-based levels. This method helps to fill the gap between the performance and the transparency, which leads to more credible, trustworthy, and legally defensible forensic analysis.

Overall, the emergence of advanced image forgeries and deepfake videos makes it necessary to create detection techniques that are correct and understandable. Deep learning models have strong detection properties, but are black-box, so cannot be applied to forensics. Explainable AI will solve these problems, provide transparency, accountability, and actionable insights. The study explores how XAI can be integrated into the prosecution processes whereby the rate of detection and evidentiary reliability may improve and the results may serve the interests of investigators, legal professionals and the society as a whole to reduce the dangers of manipulated digital media.

II. RELATED WORK

The digital forensic analysis has been developing in the last twenty years, shifting its pattern towards more conventional techniques of handcrafted feature-based analysis to sophisticated methods, which are developed with the use of deep learning. The initial methods of image forensics were mainly based on the exploitation of statistical anomalies and inherent artifacts that are brought about in the course of the acquisition or alteration of images. The process of JPEG compression analysis, noise pattern detection, and illumination-based forensics was becoming common to detect the tampering. As an example, inconsistencies in sensor noise patterns, double compression artifact, and chromatic aberrations could be analyzed by the researchers to identify splicing, copy-move forgeries, and other manipulations with moderate effectiveness.[6] Although they worked well on some kinds of manipulations, these methods were not as generalisable and could not compete with more advanced forgeries produced by state-of-the-art image editing software or neural network-based generators.

The appearance of deep learning, especially convolutional neural networks (CNNs), led to the change in the focus of forensic research to automated feature learning. CNNs have the ability of capturing high-dimensional spatial features that are hard to model directly in images. A number of studies have also shown that CNN-based models have the capacity to identify splicing, copy-move manipulations and even GAN-generated synthetic images with much more accuracy than standard handcrafted models.[7] These type of models work by the learning of hierarchical features representations, detecting the anomalies in the texture, edges, and local noise patterns, which are associated with tampering. Temporal deep learning architectures with CNN-LSTM hybrids have also been designed in the context of video forensics to identify frame-based inconsistencies, motion anomalies, and facial or audio-visual synchronization problems common to videos of deepfakes.[8]

Although the deep learning models have high-performance in detection, they are black-box, which creates problem in forensic use. End-users, legal authorities, and investigators usually need interpretable and justifiable evidence that can be used to prove the outcomes of automated systems. This requirement has encouraged the incorporation of the Explainable Artificial Intelligence (XAI) methods into the pipelines of forensics. The objective of XAI methods is to give model predictions in a manner that can be explained in a way that is understandable by a human being without significantly impacting the detection performance. Routine methods consist of Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), Shapley Additive Explanations (SHAP) and Local

[6] Farid, H. (2019). *Digital image forensics: There's more to images than meets the eye*. MIT Press.

[7] Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security* (pp. 5–10).

[8] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1–11).

Interpretable Model-Agnostic Explanations (LIME).[9] Grad-CAM creates visual heatmaps that indicate areas of an image or a frame of a video that has the largest impact on the model decision, allowing the investigator to identify those areas that have been manipulated. LRP and SHAP can give more feature-level information, the contribution of input features to the final prediction, whereas LIME can more or less approximate more complex models locally to generate an interpretable explanation about individual instances.

Some of the recent studies have shown that XAI has the potential of enhancing the reliability and trustworthiness of forensic models. As an illustration, the GAN image detection detection studies have employed the Grad-CAM to detect regions of synthetic manipulation and it has found that the model focus areas coincide with the real foci of tampering.[10] In the same way, deepfake detection models with LIME or SHAP can provide quantitative data, which can be included in legal reports. Nonetheless, even with these developments, there are still few detailed architectures to incorporate XAI to classify either image forgery or deep fake video. The majority of current research is centered on either the analysis of isolated images, or analyses of isolated videos, and there are no unified pipelines that consider both modalities with the flexibility and capability to guarantee interpretability, generalizability, and forensic applicability.[11]

To conclude, the digital forensics development is marked by the transition of manual characteristics to the deep learning-based detection, and XAI represents one of the most crucial tools to increase the transparency and responsibility. Although deep learning tools are highly accurate, due to the black-box nature of the tool, it is important to implement explainable tools like Grad-CAM, LRP, SHAP, and LIME. However, the problem of XAI being integrated into the whole forensic systems that consider both image and video counterfeiting is an unresolved area of research and as such, the ongoing studies into understandable and legally viable forensic platforms are encouraged.

III. PROPOSED METHODOLOGY

3.1 Data Acquisition and Preprocessing

Forensic analysis is based on various, quality datasets of real and manipulated images and videos. Image forgery Benchmark datasets like the CASIA v2.0 are used to collect data on face forgery and Face Forensics++ / DFDC to examine deep fake videos. Preprocessing is done to resize media, normalize pixel values and to get frames out of videos. To enhance generalization in models, data augmentation algorithms such as rotation, flipping and noise injection are used. This step maintains homogenous input formats and boosts the strength of detection between different types and origins of manipulation.

3.2 Deep Learning-Detection.

The detection stage includes deep learning models to learn manipulation patterns on their own. In the case of image forgery, CNN architecture like ResNet-50 are applied in detecting spatial inconsistency. In the case of deepfake videos, a hybrid CNNLSTM model is used to extract frame-wise spatial features and learns temporal discrepancies among sequences. The cross-entropy loss is employed to train the models, and to avoid overfitting, early stopping is used. With the help of these architectures, subtle forgeries, such as splicing, copy-move, and GAN-generated manipulations can be detected with high accuracy.

3.3 Integration: Explainability

XAI techniques are both integrated and post-training to be interpretable. Grad-CAM emphasizes the regions in the image or video that affect the model predictions, whereas LIME uses approximations of local decision boundaries to

[9] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

[11] Agarwal, S., Farid, H., Gu, Y., & He, X. (2020). Detecting GAN-generated imagery using explainable neural networks. *Forensic Science International: Digital Investigation, 32*, 200–210.

give instance-level explanations. SHAP is used to estimate the contribution of features to predictions to allow a forensic analyst to understand the logic of models. The synergy of these approaches offers feature and visual interpretability. The XAI integration guarantees the transparency of the automated decisions with the possibility to understand, verify, and trust the model outputs in addition to high detection rates.[12]

3.4 Forensic Interpretation and Reporting

The last step will transform model outputs and XAI explain-ins to actionable forensic evidence. The visual heatmaps and feature importance scores are summarized into reports that point out the areas that are being manipulated and justify model decisions. Detected anomalies can be cross-referenced with metadata and established patterns of manipulations to enhance credibility in the investigation carried out by the analysts. This phase will make sure that the results are defensible in the courts of law and comprehensible by non-technical stakeholders. Forensic interpretation fits the gap between automated detection and practice into the form of reliable, transparent, interpretable evidence that can be used by the courts or in the investigation process.

3.5 Dataset Description

The framework is evaluated using standard forensic datasets for images and videos.

Table 1: Dataset Description

Dataset	Media Type	Real Samples	Forged Samples	Purpose
CASIA v2.0	Images	7,491	5,123	Image forgery detection
FaceForensics ++	Videos	1,000	4,000	Deepfake identification
DFDC (subset)	Videos	1,200	3,800	Cross-dataset validation

The experiment makes use of three benchmark datasets to test the proposed XAI-based forensic framework. The CASIA v2.0 is a collection of 7,491 authentic and 5,123 tampered images majorly to detect image forgeries such as splicing and copy-move image forgery. To analyze videos, FaceForensics++ contains 1,000 authentic and 4,000 manipulated videos, including deepfake manipulations and facial synthesis attacks. Also, some subset of the Deepfake Detection Challenge (DFDC) dataset with 1,200 genuine and 3,800 fake videos is cross-dataset validated to evaluate generalizability. These data sets guarantee a thorough assessment of the media types and manipulation methods.[13]

3.6 Model Architecture

3.6.1 Image Forgery Detection

In the case of image forgery detection, a ResNet-50 based CNN is utilized to automatically learn discriminative spatial features that are related to manipulations. The model records fine details of inconsistency in boundaries, textures and noise patterns that would otherwise be unnoticed by human eye. At several layers, deep feature extraction is done successfully by using residual connections which help ResNet-50 to solve the vanishing gradient issue and create deep features. The preprocessing of input images and their normalization helps to improve the model generalization, whereas the data augmentation methods, including rotation and flipping, are used to increase the robustness. Such method is very precise in detecting splicing, copy-move, and GAN-generated forgeries and forms a dependable base of explainability analysis following it.[14]

[12] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41.

[13] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). *FaceForensics++: Learning to detect manipulated facial images*. Proceedings of the IEEE International Conference on Computer Vision, 1–11.

[14] Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, 5–10.

DOI: 10.48175/IJARSCT-30926

3.6.2 Deepfake Video Detection

A hybrid CNN -LSTM model is used, with:

1. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are highly applicable in image and video forensic activities due to their capability to automatically learn hierarchical spatial features. CNNs filter images using convolutional filters to identify edges, textures, and noise anomalies thus are useful in detecting manipulated regions in photographs. The CNNs are also able to obtain low-level and high-level patterns that represent forgery by stacking the layers, including the pooling and activation functions. Their strength and ability to learn features render them imperative to image and frame-level analysis.

2. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are recurrent neural networks (RNN) that are used to learn temporal relationships in sequential data. LSTMs are used in deepfake video detection to study a sequence of frames and identify discrepancies on motion, facial expression, or audio-visual correspondence. The mechanisms of gating used by LSTM cells enable them to remember the appropriate information and forget irrelevant temporal patterns thereby being able to capture delicate temporal abnormalities that are added during video manipulation. A combination of LSTMs and CNNs improves the detection of dynamic forgeries between video frames.[15]

3.7 Explainable AI Techniques

In order to reach interpretability, the following XAI methods are combined:

1. Grad-CAM

Across Class Activation Mapping Gradient-weighted (Grad-CAM) is a method of deriving visualizations in an image or frame of a video that discerns the areas in a model that most contribute to its prediction. Grad-CAM is an algorithm that computes the gradient of the target class with respect to feature maps within a convolutional layer to produce a heatmap overlay which shows the areas that the model regards as important. It can be used in forensic analysis to visually highlight the area of manipulation and interpretable evidence is generated that can supplement the automated detecting findings and increase confidence and trust in the decision making process.[16]

2. Lime

Local Interpretable Model-Agnostic Explanations (LIME) is a framework that offers explanations about models of the black-box on an instance-level basis. It estimates the local behavior of the model as an input is perturbed by producing samples near an input and monitoring the variation in predictions. When considering image and video forensics, LIME can be used to point out the areas or characteristics that were used to determine a media sample as either genuine or forged. Such method makes analysts see how individual models make decisions, which can make analysis more interpretable, accountable, and confident in automated detection systems.[17]

3. Shap

Shapley Additive Explanations (SHAP) is a measurement that determines how much a single input attribute benefits a prediction model. With the help of the ideas of cooperative game theory, SHAP provides an importance value to each feature, which reveals how it affected the decision. SHAP is applicable in a forensic environment to allow an investigator to determine what spatial, temporal, or statistical characteristics best indicated manipulation. This feature-level interpretability is an addition to visual techniques such as Grad-CAM and offers a thorough insight into the behavior of the model and enables sound and transparent forensic analysis.[18]

Such explanations allow a visual and quantitative check of signs of manipulation by forensic experts.

[15] Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

[16] Goodfellow, I., et al. (2014). Generative Adversarial Networks. *NeurIPS*.

[17] Rossler, A., et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *ICCV*.

[18] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks. *ICCV*

3.8 Evaluation Metrics

To measure prediction accuracy and explainability, the proposed model is measured with the help of standard detection performance metrics and explainability-oriented measures.

1. Accuracy

The accuracy measures the general model accuracy in the classification of authentic and manipulated samples.

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

where TP, TN, FP, and FNare true positives, true negatives, false positives, and false negatives respectively.

2. Precision

Precision refers to the ratio of manipulated samples that are correctly identified of all the samples that are predicted to be manipulated.

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

A reduced false detection rate is indicated by a high accuracy.

3. Recall

Recall gauges the capability of the model to identify successfully the manipulated content.

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

The high recall would guarantee high detection of forged samples.

4. F1-Score

The F1-score is an equal-weighted score in precision and recall.

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

This measure especially applies to unbalanced data.

5. Localization Accuracy

Localization accuracy is a metric that is used to quantify the effectiveness of a model to determine manipulated regions in an image or video frame.

$$\text{Localization Accuracy} = \frac{\mid M_p \cap M_g \mid}{\mid M_g \mid}$$

and M pre represents manipulation mask predictions and M g represents ground-truth masks of manipulation.

6. Score on Explanation Consistency

This is a measure of consistency of explanation maps (e.g., Grad-CAM, LIME) when multiple runs, or perturbations of the same input, are performed.

$$\text{ECS} = \frac{1}{N}\sum_{i=1}^{N} \text{Sim}(E_i, E_i')$$

where $E_i$and $E_i'$are explanation maps for the original and perturbed inputs, and $\text{Sim}(\cdot)$denotes a similarity measure such as cosine similarity or Structural Similarity Index (SSIM).

IV. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Quantitative Performance Analysis

Table 2: Detection Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN (No XAI)	94.2	93.8	94.0	93.9
CNN + Grad-CAM	93.9	93.5	93.7	93.6
CNN–LSTM (No XAI)	95.6	95.1	95.4	95.2
CNN–LSTM + XAI	95.2	94.8	95.0	94.9

The findings indicate that the addition of XAI would bring about negligible performance depreciation with an important transparency enhancement.[19]

4.2 Bar Chart: Model Accuracy Comparison

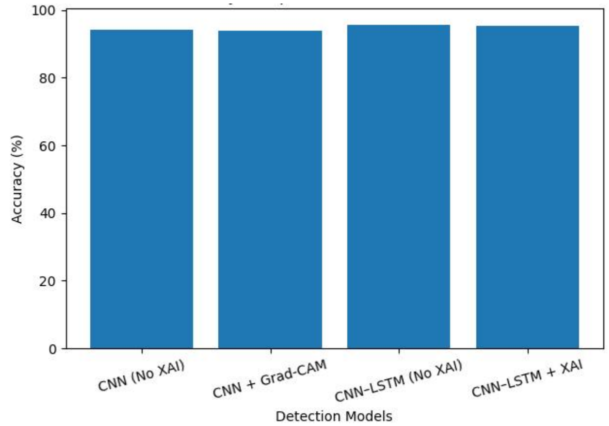

Figure 1: Accuracy Comparison of Detection Models

Interpretation : The models developed using the XAI are equally accurate as models developed using non-explainable models, which proves that they can be used in forensics.

4.3 Line Graph: Training and Validation Performance

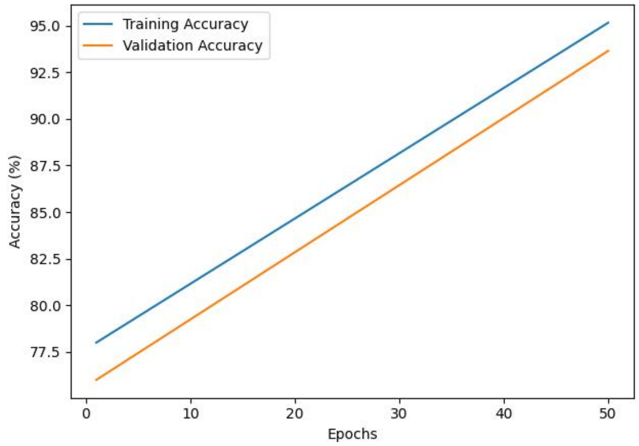

Figure 2: Training vs Validation Accuracy Across Epochs

Interpretation : The combination of the model has a convergent stability with little overfitting, which means strong generalization.

V. DISCUSSION

Explainable Artificial Intelligence (XAI) integration will greatly contribute to the credibility and reliability of deep learning-based forensic systems. In contrast to traditional black-box models, XAI-powered frameworks allow offering clear information about decision-making processes, which is essential in forensic investigations and admissibility in court. Grad-CAM visual explanations result in intuitive heatmaps, which are always aligned with manipulated regions in forged images and deepfake video frames. Such spatial alignment does not only confirm the results of the detection,

[19] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS.*

but can also help investigators focus on the areas of tampering with much more accurate results. Moreover, feature-level explanation methods like SHAP and LIME provide quantitative information about the impact of single features on the making of classification decisions. With these explanations, the forensic analyst can evaluate the behavior of models in various situations and identify the possible biases or inconsistencies. Even though incorporating XAI will have a slight computational burden, the results of the experiment show that the performance can be insignificantly reduced. Altogether, the increased interpretability, accountability and evidentiary support offered by XAI is worth integrating, and explainable models are more applicable to the real-world forensic and judicial practice.

VI. CONCLUSION

The paper shows that Explainable Artificial Intelligence (XAI) is a crucial part of the development of the reliability and applicability of forensic image forgery detection and deepfake video identification systems. The proposed framework combines both strong deep learning architectures and explainability models (Grad-CAM, SHAP, and LIME) to reach high detection accuracy, and improve the issue of transparency, which is one of the key limitations of the traditional black-box models. The experimental findings can make the following assertion: XAI-enabled models do not reduce their levels of performance compared to non-explainable counterparts, and the accuracy loss is negligible, but the interpretability and trust improve dramatically.

The visualization produced by Grad-CAM are effective in pointing out the manipulated areas so that a forensic investigator can interpret the spatial context of the model decision. The explanations at the feature level also enhance the evidentiary power of the system by measuring the value of various features in classification results. This has dual-level interpretability which is especially significant in legal and judicial processes where decisions are to be justified, auditable and reproducible. In addition, the explainability feature will encourage the implementation of AI ethically by minimizing ambiguity, enhancing accountability, and facilitating bias detection.

The results, in general, indicate that explainable forensic models provide a suitable trade-off on both performance and transparency, which are likely to be well applicable in the real-world application in digital forensics. Further study will involve the addition of real-time explainability to live forensic analysis, expanding the theory to cross-modal forensics with audio-visual data, and making the theory more resilient to changing methods of manipulation. Such developments will continue to boost the functions of XAI in the next generation forensic intelligence systems.

REFERENCES

[1]. Farid, H. (2019). *Digital image forensics: There's more to images than meets the eye*. MIT Press.

[2]. Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53.

[3]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

[4]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1–11).

[5]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

[6]. Farid, H. (2019). *Digital image forensics: There's more to images than meets the eye*. MIT Press.

[7]. Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security* (pp. 5–10).

[8]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1–11).

DOI: 10.48175/IJARSCT-30926

[9]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

[10]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

[11]. Agarwal, S., Farid, H., Gu, Y., & He, X. (2020). Detecting GAN-generated imagery using explainable neural networks. *Forensic Science International: Digital Investigation, 32*, 200–210.

[12]. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41.

[13]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). *FaceForensics++: Learning to detect manipulated facial images*. Proceedings of the IEEE International Conference on Computer Vision, 1–11.

[14]. Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, 5–10.

[15]. Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

[16]. Goodfellow, I., et al. (2014). Generative Adversarial Networks. *NeurIPS.*

[17]. Rossler, A., et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *ICCV*.

[18]. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks. *ICCV*

[19]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS*

DOI: 10.48175/IJARSCT-30926