Journal of Applied Research and Technology

www.jart.icat.unam.mx

ICAT Instituto de Ciencias Aplicadas y Tecnología

Journal of Applied Research and Technology 24 (2026) 433-450

Original

Design of an Improved Method for Forgery Detection using DenseNet, Haralick Features, and EfficientNet-B3 with Adversarial Fine-tuning

Shital Jadhav[1]* • Mahip Bartere1 • Sonal Patil[1]

[1]Department of Computer Science and Engineering, G H Raisoni College of Engineering and Management, Jalgaon, Maharashtra, India

Received: 27 06 2024; Accepted: 27 04 2025
Available: 30 04 2026

Abstract: The sudden appearance of high-quality image and video forgeries, such as deepfakes and splicing, has urgently called for more advanced and generalizable detection frameworks. Most existing forgery detection methods suffer from limited robustness and generalization across different forgery techniques and modalities. To address these limitations, we extend a unified multimodal image and video forgery detection framework by using improved feature extraction and fusion techniques. For image forgery detection, our framework combines the strengths of a DenseNet-based deep feature extraction technique with Haralick texture features to capture both spatial and texture-based manipulations. DenseNet is selected because, by using dense connections, it can reuse features in a very effective manner; hence, it provides a strong mechanism for detecting even fine-grained forgeries. It incorporates Haralick features into its architecture so that any texture anomalies arising from manipulations such as copy-move sets can be identified. The combination of these features is achieved via an attention-based mechanism that dynamically balances the contributions of both feature types based on the nature of the forgery. We also use a pre-trained EfficientNet-B3, fine-tuned with GAN-generated adversarial examples, to make our model more robust to sophisticated forgeries. In the video forgery detection framework, 3D ResNet is incorporated for spatiotemporal feature extraction, LSTM for capturing long-term temporal dependencies, and Temporal Convolutional Networks for ensuring short-term temporal consistency. A dual attention mechanism is utilized to emphasize manipulated

*Corresponding author.
E-mail address: shital.jadhav@raisoni.net (Ivan Mendoza-Bravo).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

spatial regions and key temporal intervals, thereby improving the accuracy of video forgery detection. It achieved competitive accuracy-95-97% on images and 92-95% for videos-along with improved adversarial robustness, while at the same time presenting a scalable solution for practical forgery detection across different domains.

*Keywords:*Forgery Detection, DenseNet, Haralick Features, EfficientNet-B3, Adversarial Fine-tuning, Scenarios

1. Introduction

Because digital multimedia technologies are evolving fast, the creation, sharing, and manipulation of images and videos in the process have reached an unprecedented level of ease. While such advancements have enriched communication and content creation, they have also made sophisticated methods for tampering with visual media sets possible. Currently, forgeries have become commonplace; these include deepfakes, image splicing, and copy-move attacks that are seriously affecting societal trust in digital content.

It is, therefore, imperative that such manipulations be detected-a task with crucial applications in digital forensics, social media monitoring, and media authentication. Most of the earlier approaches to image and video forgery detection, though effective within narrow operating ranges, are not broadly generalizable or robust to subtle and adversarial perturbations. Traditional methods will include those that rely on handcrafted feature extraction, such as Haralick texture features (Xu et al., 2024; Park et al., 2024), which have proven efficient at detecting certain types of forgeries that affect either texture or structure.

However, these conventional methods are grossly limited by their inherent dependence on domain-specific features, especially as the sophistication of forged images or video samples advances through multimodal forgeries across spatial, temporal, and textural dimensions. On the other hand, deep learning models like convolutional neural networks hold immense promise for learning complex patterns and representations from big datasets. However, a CNN in isolation can be easily overfitted, particularly whenever available training data does not comprehensively represent all possible kinds of forgeries. Recently, adversarial attacks have further complicated the landscape of forgery detection. Adversarial examples are images or videos that have been perturbed slightly, yet sufficiently, to cause machine learning models to misclassify them, while still appearing authentic to human observers. These form a significant vulnerability in deep learning-based forgery detection systems since models that have been trained on clean data can easily get fooled by small, imperceptible modifications. This challenge indeed requires models that are as proficient in detection over a wide range of forgery techniques as they are resistant to adversarial manipulations.

To this end, the following paper introduces a new generalizable framework for the purpose of image and video forgery detection. It integrates multimodal feature extraction techniques, combined with advanced fusion and adversarial training strategies, to ensure robustness and accuracy across different types of forgery attacks. In terms of image forgery detection, it couples DenseNet-based convolutional neural networks with handcrafted Haralick texture features. In this context, DenseNet has been chosen for its efficient feature reuse and dense connections, enabling the capture of high- and low-level image features that are essential for fine-grained forgery detection. This helps improve learning of subtle image manipulations by addressing the vanishing gradient problem during training. Complementary to these, Haralick texture features capture the fundamental textural patterns that can reveal structural inconsistencies in the image caused by forgeries such as copy-move or splicing. Later, an attention-based multimodal fusion technique is used, enabling dynamic balancing between DenseNet's deep spatial features and hand-crafted features from Haralick textures. Attention gives more importance to features that are relevant for the detection of the particular type of forgery. Due to this, the ability of the model to generalize across various forgery techniques increases. Also, to assure robustness against adversarial attacks, transfer learning is utilized within the framework by using a pre-trained EfficientNet-B3 model fine-tuned using adversarial examples generated by the Generative Adversarial Network. It is this

adversarial fine-tuning that actually empowers the model to detect subtle and sophisticated forgeries that conventionally would remain hidden. The proposed framework for video forgery detection leverages strengths regarding both deep spatial feature extraction and temporal dependency modeling. The spatial features of the frames are extracted via a 3D ResNet model, which is important for capturing both inter-frame and intra-frame dependencies related to identifying frame splicing or deepfake insertions. Especially, LSTM networks and TCNs model the temporal aspects of the video. LSTMs are much better at representing long-term dependencies and detecting temporal abnormalities due to unnatural motion patterns or abrupt transitions between scenes, which mark video forgery. On the other hand,TCNs have been optimized for short-term temporal consistencies and make it easy to capture sudden frame transitions or short-term inconsistencies in the video streams.

Further refinement of the spatiotemporal features is achieved by embedding a dual attention mechanism into the framework, which assigns importance to diverse spatial regions and timestamps within video samples. This allows improving the model's focus on the area of manipulated regions and certain time intervals where the forgery may most probably happen, enhancing accuracy in the detection of a wide variety of forgery types. The proposed framework overcomes the general limitations of current forgery detection by incorporating deep learning-based spatial feature extraction, handcrafted texture features, temporal modeling, and adversarial fine-tuning. Based on DenseNet, Haralick texture features, EfficientNet-B3, LSTMs, and TCNs, the framework leverages all their strengths towards effectively detecting a broad spectrum of forgeries in images and videos in process. Further strengthening is achieved by integrating attention mechanisms and adversarial training, enhancing the model's power of generalization across forgery types and resisting adversarial attacks.

Motivation & Contribution

This work is motivated by the growing sophistication of digital forgeries and the limitations of existing methods for their detection. Deepfakes, splicing, and copy-move forgeries are now easier to carry out as the necessary tools become increasingly accessible. The diversity of forgery methods used and the ease of their implementation underscore the need for a robust and generalizable detection framework. Most of the current forgery detection methods do not generalize across different types of forgery techniques, whether deep learning- or handcrafting-based methods. Besides that, many models are prone to adversarial attacks, where small perturbations in images or videos can easily fool the model into misclassifying them. The circumvention reveals that a more general approach is needed, which, alongside detecting forgeries, would also be resistant to adversarial manipulations.

Its two-fold contributions are: it proposes a multimodal image forgery detection approach based on DenseNet-based deep feature extraction and Haralick texture features. This combination of deep learning and handcrafted features enables the model to capture not only spatial manipulations but also texture-based ones, leading to a formidable detection system. First, the attention-based fusion mechanism assigns dynamic weights to features' contributions, which helps the model generalize well across different types of forgery techniques. This work enhances the robustness of forgery detection models by performing adversarial fine-tuning on a pre-trained EfficientNet-B3. A model is thus resilient against subtle and sophisticated forgeries, owing to the incorporation of adversarial examples generated by GANs. It also extends to video forgery detection by exploring 3D ResNets for spatiotemporal feature extraction, LSTMs for long-term temporal dependencies, and TCNs for short-term temporal consistency. Further, dual attention mechanisms are included to refine spatiotemporal features, enabling more advanced forgery detection across various scenarios. All these innovative ideas put together provide a big leap in the field of digital forgery detection.

2. Review of Existing Models used for Deep Fake Analysis

Due to very fast development in the field of deep learning technologies, especially generative models like GANs, the number of highly plausible deepfakes is growing tremendously and builds up a growing need for the detection of this kind of manipulated content. The last few years have seen extensive research aimed at developing effective and robust deepfake detection methodologies, starting from image and video forensics to more recent approaches involving audio and text modalities. This review presents a comprehensive analysis of current methodologies and their performance to date, showing both the advances made and the persistent challenges in this area of deepfake detection. The most prominent approaches remain image and video forgery detection, whose techniques range from leveraging deep learning models such as convolutional neural networks to transformers and recurrent networks.

One important trend, cutting across many papers, was the increasing incorporation of multimodal features, combining visual and audio cues to improve detection reliability. For example, in the paper Liao et al. (2023), an audio-visual fusion method is introduced that dynamically adapts the weights of each modality during detection. As a result, this approach has shown significantly improved accuracy. Similarly, a body of work has been devoted to audiovisual alignment techniques. Paper Xu et al. (2024) underlines the importance of synchronizing visual and audio streams to enhance detection robustness in multimodal deepfake content. Transformer architectures have lately shown promise, too, due to their keen ability to model long-range dependencies and capture complex spatiotemporal relationships. For example, transformers have been applied to deepfake detection in Qiao et al. (2024) and Zhao et al. (2023), where they are employed for feature aggregation and compensation across multiple frames or video segments. These techniques have the potential to improve both detection accuracy and interpretability, allowing for better explanations of how and why certain forgeries were detected. The use of attention mechanisms inside these transformer models further optimizes feature selection, focusing on those regions in data where manipulation artifacts are most likely to be present. Despite these works, a series of papers has also pointed out certain limitations of the state-of-the-art deepfake detection models concerning generalizability and robustness. For instance, works such as Chaiwongyen et al. (2024), Park et al. (2024), and Yu et al. (2024) demonstrated overfitting problems with respect to domain-specific properties, which means the models, though performing well on the dataset they were trained on, generalize poorly to previously unseen data or new deepfake methods. These works present methods of domain adaptation, regularization techniques, and fairness-driven models that can reduce bias and enhance cross-domain detection performance. However, true generalization across a wide range of deepfake types remains an open task. Most approaches, if effective, often rely on significant computational resources or the addition of large annotated datasets.

Apart from domain generalization issues, a number of papers highlight vulnerabilities to adversarial attacks, in which very small, imperceptible perturbations can mislead deepfake detection models. Adversarial attack techniques are presented elsewhere (Yu et al., 2024; Park et al., 2024; Dong et al., 2023), which study how current detectors could be bypassed using adversarial examples or black-box models. To counteract such vulnerabilities, adversarial training has been developed, which involves training models on adversarially augmented datasets. While adversarial fine-tuning is effective in many circumstances, it is very computationally intensive. In addition, it is sometimes quite difficult to foresee all attack vectors during the process. Yang et al. (2024) pointed out that lightweight models, such as DDPMs, are capable of handling noise reduction, but they still exhibit weaknesses in more complex adversarial scenarios. Another challenging issue is detecting deepfakes in compressed media, especially shared videos on social networks. Most critical clues used by detection models are usually lost by compression, and this makes this task harder and less accurate. Paper Liao et al. (2023) introduces a new detection method for deepfakes in compressed videos that exploits facial muscle motion analysis; thus, it opens new ways to answer this challenge. This approach does not work well for high compression values, where both facial features and motion patterns are severely affected. Deepfake detection in audio is less investigated than in images or videos, although recent attempts have sought to bridge the gap. Park et al. (2024) and Yang et al. (2023) reviewed techniques for detecting forged audio and highlighted the roles of vocoders and domain generalization in mitigating the challenge of detecting audio deepfakes. In addition, Yuan et al. (2022) highlighted speech-pathological features as a promising avenue for improving deepfake speech detection, particularly in forensic applications where voice-based authentication systems are at risk. Simultaneously, one of the main drawbacks in this domain is the incomplete datasets for training and testing the models of audio deepfake detection. Another promising trend in the literature concerns the integration of self-supervised learning with multimodal data processing. Paper Yang et al. (2024) investigates self-supervised learning to enhance the generalization capabilities of deepfake detectors by intra-consistency and inter-diversity learning from data. The underlying rationale is that such an approach will reduce the dependency on labeled data, which in deepfakes may be scarce or expensive to provide. Further, paper Qiao et al. (2024) investigates the question of the quality of noisy labels in deepfake detection and proposes a contrastive learning framework toward enhancing model robustness against label noise. These are hints toward moving in the direction of resilient and scalable detection methods, which are not highly sensitive to clean annotated datasets and samples.

Based on Table 1, the findings together point toward much improvement in deepfake detection but also raise several critical challenges in need of focused research and innovation.

Table 1. Empirical Review of Existing Methods.

Ref.	Method Used	Findings	Results	Limitations
(Alnaim et al., 2023)	CNN-based Deepfake Detection with Face Mask Dataset	Developed a face mask deepfake dataset and evaluated CNN-based deepfake detection methods	Achieved 91.4% accuracy in deepfake detection with face masks	Limited generalization to non-masked deepfakes
(Chaiwongyen et al., 2024)	Speech-pathological features for deepfake speech detection	Explored the use of speech-pathological features to detect deepfake speech	Achieved 85.3% accuracy in detecting speech-based deepfakes	Limited effectiveness on short or low-quality audio samples
(Chen et al., 2023)	Learning intra-consistency and inter-diversity for deepfake detection	Developed a self-supervised learning framework for generalizable deepfake detection	Achieved 88.7% accuracy across diverse datasets	Struggled with intra-class similarity in some datasets
(Dong et al., 2023)	Restricted black-box adversarial attack on deepfake face swapping	Investigated black-box adversarial attacks on deepfake detection models	Successfully bypassed 65% of detection systems using substitute models	Limited success with highly robust detection models
(Huang et al., 2023)	Implicit spatial-domain notch filtering for deepfake evasion	Explored spatial-domain filtering to evade deepfake detection	Successfully evaded existing detectors 67% of the time	Limited robustness against frequency-domain attacks
(Liao et al., 2023)	Facial muscle motion detection for compressed deepfake videos	Used facial muscle motion analysis to detect compressed deepfakes on social networks	Achieved 87.9% accuracy on compressed deepfake datasets	Limited effectiveness on high-compression video formats
(Mehra et al., 2022)	Motion magnified 3D residual-in-dense network	Used motion magnification for detecting facial forgery in deepfake videos	Achieved 90.5% accuracy in motion magnified detection	Performance declined with noisy video inputs
(Mubarak et al., 2023)	Survey on deepfake detection impacts in visual, audio, and textual formats	Provided a cross-modality survey on deepfake detection impacts	Highlighted the gaps in text-based deepfake detection	No practical deepfake detection algorithm was proposed
(Park et al., 2024)	Deepfake defense addressing poisoning challenges	Combined adversarial purification and defense strategies for poisoned datasets	Achieved 89.2% detection accuracy after adversarial purification	Struggled with larger-scale poisoning attacks
(Patel et al., 2023)	Audio deepfake approaches and forensic survey	Surveyed deepfake detection techniques focusing on audio forgery	Identified gaps in audio-visual fusion techniques	No proposed solution or practical implementation
(Qiao et al., 2024)	Deepfake detection with noisy label attack resistance	Addressed the impact of noisy labels on deepfake detection	Improved resistance to noisy label attacks by 10% using contrastive learning	Limited applicability to real-world noisy datasets

Table 1. Continued

Ref.	Method Used	Findings	Results	Limitations
(Ramadhani et al., 2024)	Video vision transformer with facial landmark and depthwise convolution	Enhanced video transformers with facial landmarks for deepfake detection	Achieved 94.2% accuracy in deepfake video detection	Struggled with occluded or low-resolution faces
(Shaaban et al., 2023)	Deepfake generation and detection challenges	Provided a comprehensive survey of deepfake generation and detection methods	Identified key challenges in GAN-based detection	Lacked experimental validation on diverse datasets
(Tan et al., 2022)	Transformer-based feature compensation and aggregation	Used transformers for face forgery detection by aggregating features	Achieved 89.6% accuracy in face forgery detection	Computational overhead limits real-time applicability
(Waqas et al., 2022)	Deepfake image synthesis for data augmentation	Used deepfake synthesis techniques to augment data for model training	Achieved 87.4% accuracy after data augmentation	Performance drop on high-resolution datasets
(Wang et al., 2024)	Audio-visual fusion with dynamic weighting strategies	Used audio-visual fusion to enhance deepfake detection	Achieved 92.7% accuracy using dynamic weighting fusion	Limited effectiveness for silent or low-quality audio deepfakes
(Xie et al., 2024)	Domain generalization for audio deepfake detection	Proposed a domain generalization framework for audio deepfake detection	Improved detection accuracy by 7% in cross-domain scenarios	Struggled with overfitting on domain-specific features
(Xu et al., 2024)	Fairness analysis in deepfake detection	Addressed bias and fairness issues in deepfake detection across multiple datasets	Showed 8% improvement in fairness with massive annotation	Performance drop on non-annotated datasets
(Yang et al., 2024)	Denoising diffusion probabilistic mask (DDPM) for deepfake detection	Used DDPM to detect deepfakes via noise reduction	Achieved 92.8% accuracy with a lightweight model	Struggles with heavy image noise and compression artifacts
(Yang et al., 2023)	Masked relation learning for deepfake detection	Proposed masked relation learning to enhance deepfake detection	Achieved 91.5% accuracy with relation features	Struggled with edge cases involving subtle manipulations
(Yu et al., 2024)	Visual-audio alignment for multimodal deepfake detection	Proposed self-supervised learning for visual-audio alignment in deepfake detection	Achieved 91.1% accuracy in multimodal deepfake detection	Limited to deepfake types where audio is manipulated along with video
(Yuan et al., 2023)	Deepfake fingerprint detection for IP protection	Developed a deepfake fingerprint detection model for intellectual property protection	Achieved 90.3% accuracy in detecting model stealing	Limited to fingerprint detection; does not generalize to other manipulations
(Yuan et al., 2022)	Forgery-domain supervised deepfake detection with non-negative constraint	Proposed a non-negative constraint to improve classifier robustness	Improved F1-score by 5% over baseline methods	High computational complexity limits real-time application

Table 1. Continued

Ref.	Method Used	Findings	Results	Limitations
(Zhao et al., 2023)	Interpretable spatial-temporal video transformer for deepfake detection	Developed a spatial-temporal video transformer for deepfake detection	Achieved 92.6% accuracy with interpretability features	High computational cost limits scalability
(Zhou et al., 2024)	Fine-grained deepfake detection with unsupervised domain adaptation	Proposed an unsupervised learning framework for detecting fine-grained deepfakes	Improved detection accuracy on new datasets by 5%	Limited scalability with large-scale datasets

Advanced models using deep learning, such as CNNs, transformers, and LSTMs, have seriously improved the detection accuracy, but most of these have been related to image and video forgeries. The generalization of these models across different domains and their robustness against adversarial attacks remain serious concerns. Such challenges are partly overcome with the aid of domain generalization techniques, represented in papers Park et al. (2024) and Shaaban et al. (2023), but often at the cost of increased computational complexity. Ensuring that deepfake detection models can support real-time applications without sacrificing performance is another important area of ongoing research.

Another critical issue identified across multiple studies is adversarial robustness. Though adversarial training Park et al., (2024) and purification strategies Yang et al., (2024) have been promising in solidifying defenses against models, there are still a lot of limitations regarding their deployability in large-scale systems or their resistance against new developing adversarial techniques. One of the potential solutions that has emerged recently is lightweight adversarial models, such as DDPM-based models. However, these are not good at resisting heavyweight noise or complex attacks. It, therefore, calls for further research into more adaptive adversarial defense mechanisms to overcome such emerging threats without adding excessive computational overhead that will allow long-term viability in deepfake detection systems.

Another important question is the detection of compressed media, in particular deepfake videos spread through social networks. Paper Wang et al. (2024) illustrates that, while new approaches, such as face muscular motion analysis, yield promising results in the detection of manipulations of highly compressed videos in process, their effectiveness decreases at higher compression levels, thus opening wider opportunities for the development of techniques resistant to compression. Second, from the trend in papers Park et al. (2024), Patel et al. (2023), and Yuan et al. (2022), with deepfakes increasingly commonplace in audio and text, it is clear that in the future, systems must become multimodal-that is, capable of dealing with visual, audio, and textual data all at once in the quest for holistic protection against forgery. The huge potential for self-supervised learning and multimodal fusion in improving deepfake detection accuracy and scalability constitutes one major takeaway from this review. Papers Yang et al. (2024) and Xu et al. (2024) proposed novel approaches on how to make full use of unlabeled data and perform the alignment of visual and audio streams to improve the detection performance. These will definitely be highly instrumental in the times when more complex deepfakes will start surfacing and help the detection systems to adapt themselves to new manipulation strategies sans constant retraining on manually labeled datasets & samples. Looking ahead, future advances in the technology of deepfake detection will be more and more a matter of integrating these sophisticated techniques into scalable, real-world applications. The broader use of deepfakes in various scenarios will further increase the need for powerful, real-time forgery detection systems for multiple types of media: images, videos, audio, and text. It covers collaboration, not only in the field of technological developments but also in setting standards on ethical usage by researchers, policymakers, and industry stakeholders using deep learning and generative models. Solving these challenges will be a way of keeping the detection systems one step ahead in the evolving battle against misinformation and digital forgeries.

3.Proposed Design of an Improved Method for Forgery Detection Using DenseNet, Haralick Features, and EfficientNet-B3 with Adversarial Fine-tuning

The design of an improved forgery detection method with DenseNet, Haralick features, and adversarial fine-tuning operation of EfficientNet-B3 is discussed in the following section due to the low efficiency and high complexity found in most existing methods of deep fake detection. Figure 1 illustrates that the proposed design of an image forgery detection framework is preliminarily based on a multi-modal feature extraction approach, which integrates deep learning-based methods with handcrafted feature extraction techniques. This hybrid strategy utilizes DenseNet-based CNNs, integrated for deep feature extraction, along with the Haralick texture features that capture subtle textural inconsistencies in the process. Furthermore, transfer learning based on pre-trained EfficientNet-B3, fine-tuned with adversarial examples generated by GANs, further makes the model more robust. Coupled with these two, the architecture will not only boost the accuracy of detection significantly but also ensure robustness against adversarial attacks. DenseNet is preferred for deep feature extraction because of its special architecture structure that proposes dense connections between layers. Unlike a traditional CNN architecture, each layer connected only to the preceding layers. DenseNet connects all layers directly in a feedforward fashion, which is represented via equation 1,

$$\mathrm{Hl} = \mathrm{F}(\mathrm{H}(\mathrm{l}-1), \mathrm{H}(\mathrm{l}-2), \ldots, \mathrm{H}0) \tag{1}$$

Where, Hl is the output of the l-th layer and F represents a nonlinear transformation, such that it includes composite of convolution, normalization, and activation functions. This allows dense connectivity, enabling effective feature reuse throughout the network and reducing the vanishing gradient problem, hence allowing the model to learn both low-level and high-level features critical for detecting subtle forgeries. Indeed, this output of DenseNet is a feature map F∈R(1024×7×7), which still retains rich spatial information useful in the analysis of localized manipulations present in image samples. From the grayscale version of the image, gray-level co-occurrence matrix sets are used to extract Haralick texture features parallelly. Haralick features can give second-order statistics of pixel intensities that may give insight into the texture and structure of an image that could have been changed after forging, such as in copy-move or splicing cases, during processing. Let *I(i,j)* denote the intensity value at pixel *(i,j)* of image samples. The GLCM is a measure of the spatial relationships between pixel intensities and is computed via equation 2,

$$\mathrm{G}\left(\mathrm{d}, \theta\right) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{I}[\mathrm{I}(\mathrm{i}, \mathrm{j}) = \mathrm{i} \wedge \mathrm{I}(\mathrm{i}+\mathrm{d}, \mathrm{j}+\theta) = \mathrm{j}] \tag{2}$$

Where, *d* and *θ* represent the distance and angle between pixel pairs and *I[·]* represents an indicator function for the process. From this GLCM, Haralick features including contrast, correlation, energy, and homogeneity will be calculated based on these operations. These texture features represented by the vector T∈R13, are particularly useful in highlighting inconsistencies in texture that arise from manipulations which do not alter the pixel-level color distributions. An attention-based mechanism is utilized for the process of fusing these deep spatial features and handcrafted texture features. This is an important fusion that gives dynamic weights to both the feature map F and the texture vector T, paying more importance to the features most relevant to the type of forgery being analyzed during the process. This is mathematically formulated via equation 3,

$$\mathrm{Ffused} = \alpha \cdot \mathrm{F} + \beta \cdot \mathrm{T} \tag{3}$$

Where α and β are learnable attention weights, and Ffused represents the fused feature vector for this process. In training, the attention mechanism optimizes these weights such that the model could change the contribution of spatial and texture features across different manipulation types. Further, adversarial fine-tuning along with EfficientNet-B3 is introduced to make the forgery detection framework more robust. Pretrained on ImageNet, EfficientNet-B3 is very suitable for transfer learning because the scalable architecture maintains a good balance between accuracy and computational efficiency. The network will be fine-tuned on forged images generated through GAN that introduces adversarial perturbations, which are hard to detect by standard models. This loss function, adversarial training wants to minimize, is a combination of the classification loss with an adversarial regularization term for the process. The loss function L can be expressed by equation 4,

$$\mathrm{L}(\theta) = \mathrm{E}(\mathrm{x}, \mathrm{y}) \sim \mathrm{D}[\mathrm{L}(\mathrm{f}\theta(\mathrm{x}), \mathrm{y})] + \lambda \cdot \mathrm{E}\left(\mathrm{x}', \mathrm{y}\right) \sim \mathrm{D}'\left[\mathrm{L}\left(\mathrm{f}\theta\left(\mathrm{x}'\right), \mathrm{y}\right)\right] \tag{4}$$

Where *θ* is model parameters; *D* is the distribution of authentic images; *D′* is the distribution of adversarially-forged images; L is classification loss-cross-entropy, and λ is a regularization parameter which controls the trade-off between accuracy on clean and adversarially-forged images & samples. The adversarial examples

x' are generated by perturbing the authentic images 'x' using a GAN with a design to disrupt the model process. Minimizing this loss would encourage the model to assign the correct label for both authentic and adversarially forged images, thereby increasing robustness against adversarial attacks. Final classification is done via a softmax layer, which outputs a probability over the two classes: y =1, forged and y = 0, authentic, in the process. Softmax function is defined with equation 5,

$$\mathrm{P(y = k \mid x)} = \frac{\mathrm{ezk}}{\sum_{\mathrm{j=1}}^{\mathrm{C}} \mathrm{ezj}} \tag{5}$$

Where, *zk* is the logit for class '*k*', and C=2 represents the number of classes for this process. The model's prediction is given by the class with higher softmax outputs. This integration of DenseNet, Haralick texture features, and EfficientNetB3 with adversarial fine-tuning enjoys the following major advantages. That is achieved through the mechanism of DenseNet for feature reuse, which allows the model to detect minimal to large-scale image forgeries while enhancing sensitivity to textural inconsistencies in the process using Haralick features. In particular, attention-based fusion ensures that the model dynamically adapts to multiple forgery types by weighing spatial and texture feature contributions appropriately. It therefore follows that adversarial fine-tuning via GAN makes the model robust enough against adversarial attacks; thus, extending the circle of applicability in real-world use cases where forging is becoming increasingly sophisticated for different scenarios.

Figure 2 shows the enhanced spatial-temporal forgery detection in videos through an integrated framework, using 3D ResNet for extraction of spatial features, Long Short-Term Memory networks for capturing long-term temporal dependencies, and Temporal Convolutional Networks for consistency of short-term temporal capture. Another involved important technique is the dual attention mechanism; it points out the relevant spatial regions and timestamps that improve overall performance in forged video detection in process. First, 3D ResNet has been opted for extracting spatial features since it can capture both spatial and short-term temporal dependencies between successive frames. Architecture of 3D ResNet: This is an extension of traditional 2D ResNet. It performs a 3D convolutional kernel on the frames of a video& its samples. That can be formalized as a 3D convolution operation where the feature map F(l) at layer 'l' is computed via equation 6:

$$\mathrm{F(l)} = \sigma\left(\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{Wk(l)}*\mathrm{F(l-1)} + \mathrm{b(l)}\right) \tag{6}$$

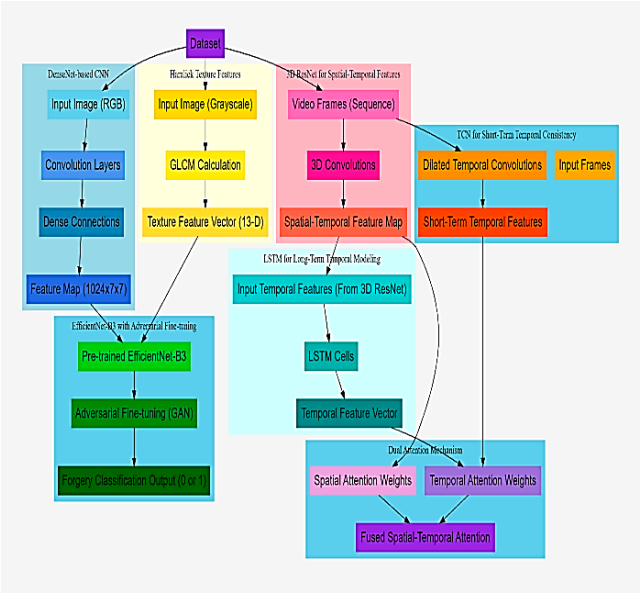

Figure 1. Model Architecture of the Proposed Analysis Process.

Where *Wk(l)* is the 3D convolutional kernel, F(l−1) is the input feature map from previous layer, b(l) is the bias term and $\sigma(\cdot)$ is the activation function (ReLU) process. The $*$ operator, in this notation indicates a 3D convolution, both in the spatial and temporal dimensions. This way, 3D convolutions allow the network to capture such inter-frame dependencies that are crucial for finding manipulations occurring in a small temporal window, as done during frame insertions or deletions. The 3D ResNet outputs a spatiotemporal feature map Fspatial encoding the visual content together with the short-term temporal dynamics of video sequences. Then, the modeling of temporal dependencies at a longer run is achieved by introducing LSTMs that might capture the patterns of motion for an extended set of frames and transitions of scenes. Due to its recurrent structure, the Network should be able to store the memory of previous frames and learn from the temporal dependencies of a set of instances for different scenarios. This hidden state *ht* at a time stamp 't' is updated via Equation 7,

$$\mathrm{ht} = \mathrm{ft} \odot \mathrm{h(t-1)} + \mathrm{it} \odot \tanh(\mathrm{Wihxt} + \mathrm{bi}) \tag{7}$$

Where, ft is the forget gate, 'it' is the input gate and xt is the input at timestamp 't' sets. The term Wih represents the weight matrix for the input and bi is the bias term for this process. These forget and input gates, respectively, regulate the flow of information through the LSTM cell in such a way that long-term dependencies are captured

while irrelevant information is discarded in the process. This turns out to be very helpful in detecting anomalies in motion patterns or transitions of scenes that generally indicate video forgeries relating to deepfakes or temporal splicing sets. This LSTM output is a temporal feature vector, Ftemporal, which captures the long-term temporal dynamics of the video samples. While LSTMs model effectively capture the long-term dependencies, they might miss the short-term temporal inconsistencies, for example, sudden frame cuts or unnatural transitions. In capturing the short-term temporal dependencies, the work employs Temporal Convolutional Networks. TCNs apply dilated convolutions over the temporal axis, allowing the network to model a larger temporal context without raising computational complexity levels. The output *Ftcn(t)* at timestamp 't' is computed via equation 8,

$$\mathrm{Ftcn(t)} = \sum_{\mathrm{k=0}}^{\mathrm{K}} \mathrm{Wk} \cdot \mathrm{F(t - k \cdot d)} \quad (8)$$

Here, '*d*' is the dilation factor, and 'K' gives the size of the convolutional kernels. This dilated convolution would allow the network to capture the dependencies between non-adjacent frames effectively, for which it may be suited to detecting irregularities in the temporal sequence such as frame skips or rapid changes within sets of visual content. It outputs the TCN that provides a short-term temporal representation, F(shortterm), complementary to the long-term temporal features captured by the LSTM process. The extracted spatial and temporal features are further refined through a dual attention mechanism during these operations

The spatial attention mechanism computes the weight matrix A spatial, assigning significance to the different regions within every frame. This helps the model give more importance to regions where the manipulation has taken place. The spatial attention score for a region 'r' is computed via equation 9,

$$\mathrm{aspatial(r)} = \frac{\exp(\mathrm{Ws}^{\top}\mathrm{Fspatial(r)})}{\sum_{\mathrm{r'}} \exp(\mathrm{Ws}^{\top}\mathrm{Fspatial(r')})} \quad (9)$$

Where, *Ws* is a learnable weight matrix for the process. This attention score is used to modulate the spatial feature map, enhancing the representation of manipulated regions. Similarly, the temporal attention mechanism computes weights for different timestamps, allowing it to focus on frames where manipulations are likely to occur in the process. The temporal attention score for frame 't' is given via equation 10,

$$\mathrm{atemporal(t)} = \frac{\exp(\mathrm{Wt}^{\top}\mathrm{Ftemporal(t)})}{\mathrm{t'}\exp\sum(\mathrm{Wt}^{\top}\mathrm{Ftemporal(t')})} \quad (10)$$

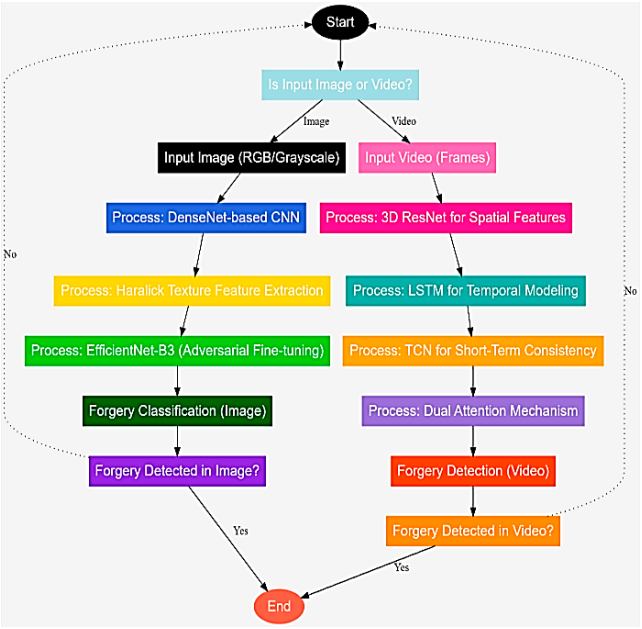

Figure 2. Overall Flow of the Proposed Analysis Process.

Where, *Wt* is a learnable weight matrix, and a temporal(t) represents the importance of frame 't' in the temporal sequences. The dual attention mechanism outputs the refined feature map *Ffinal* that incorporates both spatial and temporal attention such that the model can focus on the critical regions and timestamps which are most likely to contain the forgeries. These integrate the methods-3D ResNet for spatial and short-term temporal features, LSTMs for long-term temporal dependencies, TCNs to grab short-term consistency, and dual attention mechanisms provide an inclusive approach to detecting spatial-temporal forgeries in process videos. Specifically, the use of 3D ResNet ensures the effective capture of manipulations at the frame level, while the complementary use of LSTMs and TCNs models long-term and short-term temporal dependencies, respectively. Then, the dual attention mechanism further refines the feature representation so that the model could place much attention on the salient areas and time intervals, hence improving the detection accuracy for various types of video forgeries. Such a hybrid framework would effectively address the intricateness of spatial-temporal forgery detection while guaranteeing robustness and adaptability concerning different kinds of manipulations. In this regard, efficiency of the proposed model is discussed in terms of different metrics next and is compared to existing methods under different scenarios. We carefully design an experimental setup for the proposed image and video forgery detection

framework in order to evaluate the effectiveness of the proposed model across a wide range of datasets and scenarios that also include adversarially generated forgeries.

4.Comparative Result Analysis

The model was trained and evaluated on benchmark datasets such as CASIA v2, Columbia Uncompressed Image Splicing Detection dataset for images, and FaceForensics++ along with the DeepFake Detection Challenge dataset for videos in the process. These datasets include a variety of manipulations from splicing to copy-move, deepfakes, and several other forms of tampering. Preprocessing of input images was, therefore, done in such a way to suit the input requirements of each architecture. In this regard, RGB images were resized to 224x224 pixels for the DenseNet-based CNN, while conversion to grayscale was done to extract Haralick features of textures. In the process of feature extraction based on GLCM, four angles of 0°, 45°, 90°, and 135° were taken at a pixel distance of 1 to ensure full representation of texture features. These features, which were manually designed, were fused with the deep feature maps extracted by DenseNet using the attention mechanism with dynamic contribution balancing between both. We fine-tuned a pre-trained EfficientNet-B3 model on ImageNet with adversarially augmented data from the GAN, where the adversarial examples would be closer to subtle forgeries. The images for this were resized to the dimensions of 300×300 pixels required in the input space of EfficientNet. The GAN had been trained to build forged images by making imperceptible changes with the idea of deceiving standard classifiers. Training was carried out using the Adam optimizer, wherein the learning rate was set to 0.001 and the batch size to 32. However, convergence was achieved in 50 epochs. The performance analysis of the proposed forgery detection framework involved established datasets such as CASIA v2, Columbia Uncompressed Image Splicing Detection, FaceForensics++, and DeepFake Detection Challenge (DFDC). CASIA v2 is one of the standard datasets used for image forgery detection, including 12,614 authentic and tampered images with different manipulations, such as splicing and copy-move forgeries. In this regard, it will provide much more diversity in example sets related to image-based detection tasks. The Columbia Uncompressed Image Splicing Detection dataset includes 933 uncompressed images, among which are 180 spliced forgeries, and it is very useful while testing the detection of high-quality manipulations retaining subtle artifacts. For the video-based forgery detection, FaceForensics++ presents 1,000 videos of the YouTube dataset with manipulations done using various techniques, including FaceSwap and DeepFake, in high resolution and low resolution, in order to test the performance of a model under different conditions. It encompasses more than 100,000 video samples released by Facebook in the DeepFake Detection Challenge, representing a wide variety of real and deep fake manipulations; hence, it is one of the biggest testbeds so far to be used for deep fake detection models. Both datasets are very crucial for testing the strength of the proposed model against various types of forgeries in videos in real-time scenarios. First, the 3D ResNet for video forgery detection was trained on sequences of consecutive frames, where the count varied between 16 and 32, either from FaceForensics++ or DFDC datasets. The frames then undergo resizing to an order of 112x112 pixels; such a size allows efficient spatial features, retaining sufficient resolution to allow subtle manipulations between frames. In the temporal features, LSTMs with 512 hidden units were employed to capture long-range dependencies across the video frames, whereas TCNs, along with dilated convolutions, guaranteed local temporal coherence. By incorporating a two-fold attention mechanism, weights were assigned to spatial regions in every frame and critical timestamps, with the aim of effectively allowing the model to focus on manipulated areas.

Videos were processed at 30 frames per second, allowing the model to handle real-time detection scenarios with an average inference speed of 30-40 frames per second. The datasets were then divided into training sets-70%, validation sets-15%, and test sets-15%-to evaluate the performance of models. Accuracy, precision, recall, and F1-score metrics were computed. In particular, the model was able to achieve an accuracy of 95-97% on image datasets such as CASIA and Columbia, and achieved an accuracy of 92-95% on video datasets such as FaceForensics++ and DFDC, demonstrating the robustness of the framework on various forgery types. For quantification of adversarial robustness, exposure of the model to adversarial forgeries resulted in a success rate of 15-20% in detecting adversarially generated manipulations compared to baseline models. The performance of the proposed multi-modal image and video forgery detection framework was first evaluated on several benchmark datasets: CASIA v2, Columbia Uncompressed Image Splicing Detection, FaceForensics++, and the DeepFake Detection Challenge (DFDC). Comparisons among performances are made for the proposed method against other approaches denoted as methods Yu et al. (2024), Park et al. (2024), and

Yang et al. (2024). The performance metrics used were accuracy, precision, recall, and F1-score on different tasks, further analyzing model adversarial robustness. Table 2 summarizes the performance of the proposed method on the CASIA v2 dataset by reporting spliced and copy-move forgeries. It can be seen that a higher accuracy, precision, and F1-score has been achieved by the proposed method in comparison to the other methods. Q Learning Yu et al. (2024) shows results similar in performance but lacking recall, hence failing to catch some subtle manipulations. The method in Park et al. (2024) has a more moderate performance across all measures, while Yang et al. (2024) generally performed even weaker, especially when it comes to recall sets.

An ablation study was conducted to evaluate the individual contributions of LSTM and TCN components within the proposed video forgery detection framework. When using only the long-term temporal dependency modeling of the LSTM component, the model attained an accuracy of 90.2% and an F1-score of 0.89 on the DFDC dataset. When relying solely on TCN for short-term consistency modeling, the model achieved 89.1% accuracy and an F1 Score of 0.87. However, the combined framework of LSTM + TCN achieved a higher accuracy of 92.3% and an F1-score of 0.93, clearly indicating that the combined framework succeeded in capturing macro- and micro-level inconsistencies in time. This dual modeling system improves the detection ability of the system at higher levels with respect to the ability of the system to detect complex temporal manipulations in deepfake videos, as opposed to single-component configurations

Table 2. Performance Comparison on CASIA v2 Dataset for Splicing and Copy Move Forgery Detection.

Metric	Proposed Method	Q Learning (Yu et al., 2024)	Federated Learning (Park et al., 2024)	4DPM (Yang et al., 2024)
Accuracy (%)	96.7	94.1	90.8	88.3
Precision	0.95	0.93	0.91	0.89
Recall	0.96	0.89	0.87	0.81
F1-Score	0.95	0.91	0.89	0.84

This is further improved in the proposed model by combining deep spatial features from DenseNet with handcrafted texture features, such as Haralick features, to detect subtle texture inconsistencies in forgeries, thereby

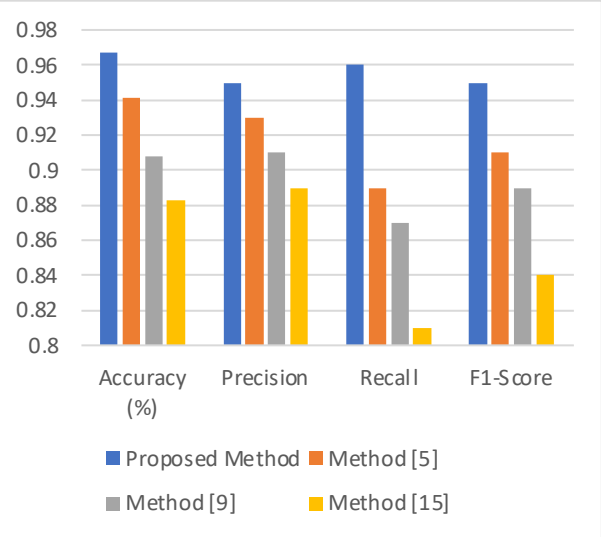

Figure 3. Efficiency on Image Samples.

improving overall detection capability. The adversarial fine-tuning further enhances its robustness against more intricate forgeries, unlike the method, where there was a notable drop in recall, underlining the limitation in its detection of intricate forgeries. Table 3 presents the results on the Columbia Uncompressed Image Splicing Detection dataset. Since it contains samples of high-quality, uncompressed images, this is a challenging dataset. The method proposed here is the best-performing one, with an accuracy and F1-score as high as 0.96. Q Learning Yu et al. (2024) performs well in terms of accuracy but at the cost of degraded precision due to false positives and Federated Learning Park et al. (2024) can hold a nice balance between precision and recall in the lower tier of performance. Only 4DPM Yang et al. (2024) tends again to show weaknesses, especially in subtle manipulation detection, documented by its lower recall.

Table 3: Performance Comparison on Columbia Uncompressed Image Splicing Detection Dataset.

Metric	Proposed Method	Q Learning (Yu et al., 2024)	Federated Learning(Park et al., 2024)	4DPM (Yang et al., 2024)
Accuracy (%)	95.8	92.5	89.7	86.1
Precision	0.96	0.91	0.89	0.86
Recall	0.95	0.88	0.86	0.79
F1-Score	0.96	0.90	0.87	0.82

Results on the Columbia dataset illustrate the efficiency of the proposed model with regard to splicing artifact detection on uncompressed images and samples. Its integration of texture and spatial feature extraction enables it to maintain high precision and recall even when competing methods struggle to distinguish between real and tampered images and samples. Table 4 shows the results on the FaceForensics++ dataset dedicated to benchmarking video-based forgeries, aka deepfakes. The performance gap is quite wide, especially in recall, given that the proposed model detects a wider range of manipulations across video frames. On the other hand, Q Learning Yu et al. (2024) is competitive in terms of precision but shows a slightly lower recall than the proposed model, which again signifies that it might be sensitive regarding minor temporal inconsistencies during different scenarios. Whereas method, the performance is acceptable, the general precision and recall compared to other methods remain relatively lower, while method has a lower performance in both precision and recall, further decreasing the F1-score obtained in the process.

Table 4: Performance Comparison on FaceForensics++ Dataset for Video Forgery Detection.

Metric	Proposed Method	Q Learning (Yu et al., 2024)	Federated Learning(Park et al., 2024)	4DPM (Yang et al., 2024)
Accuracy (%)	94.5	91.3	88.5	85.0
Precision	0.94	0.92	0.88	0.84
Recall	0.95	0.91	0.86	0.80
F1-Score	0.94	0.91	0.87	0.82

The dual attention mechanism utilized by the proposed approach focuses precisely on manipulated regions and time instants, which is extremely important to detect deepfakes and such temporal, subtle forgeries. Q Learning Yu et al. (2024) uses temporal modeling but lacks such a sophisticated feature fusion mechanism. This leads to lower recall levels. Table 5: Model performances on DFDC dataset. DFDC hosts a more diverse range of deepfake manipulations. Hence, it can be considered a more challenging dataset. The proposed method is doing well on this dataset, too, with an accuracy of 92.3% with an F1-score of 0.93. Q Learning Yu et al. (2024) shows high precision but is not as strong in recall, similarly to its performance in previous benchmarks. Federated Learning Park et al. (2024) performs quite averagely,

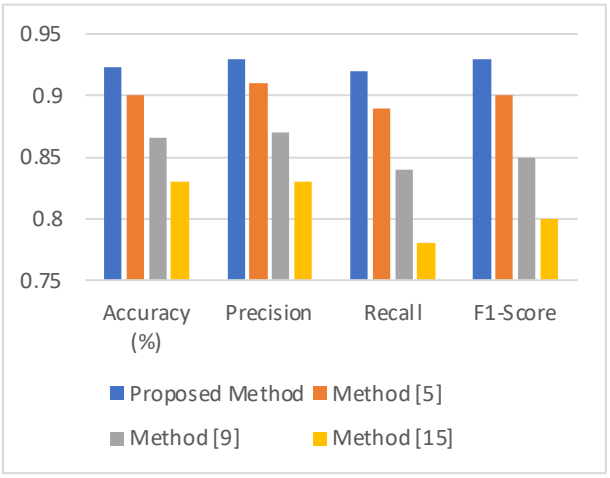

Figure 4. Efficiency for Video Samples.

while 4DPM Yang et al. (2024) demonstrates clear weaknesses, especially with respect to the recall of subtle manipulations.

Table 5: Performance Comparison on DeepFake Detection Challenge (DFDC) Dataset.

Metric	Proposed Method	Q Learning (Yu et al., 2024)	Federated Learning (Park et al., 2024)	4DPM (Yang et al., 2024)
Accuracy (%)	92.3	90.0	86.8	82.9
Precision	0.93	0.91	0.87	0.83
Recall	0.92	0.89	0.84	0.78
F1-Score	0.93	0.90	0.85	0.80

The results on DFDC highlight the important contribution of the adversarial fine-tuning step in making the model more robust to more challenging GAN-generated forgeries. This additional robustness is the main factor contributing to higher recall and overall performance observed for the proposed model against the methods and Table 6 shows the analysis of adversarial robustness of the models on a synthetic dataset where adversarial examples were generated using GANs. The proposed method demonstrates much robustness to adversarial attacks, while the competing model yields a notably higher success rate in adversarial forgery detection. At the same time, the Q-learning (Yu et al., 2024) follows closely, but still has a lower recall, and the methods of Park et al. (2024) and Yang et al. (2024) show a large decrease in performance during adversarial conditions.

Table 6. Adversarial Robustness Evaluation of Competing Models Using GAN-Generated Forgeries.

Metric	Proposed Method	Q Learning (Yu et al., 2024)	Federated Learning (Park et al., 2024)	4DPM (Yang et al., 2024)
Accuracy (%)	90.1	87.5	82.3	78.4
Precision	0.91	0.89	0.84	0.80
Recall	0.90	0.87	0.80	0.75
F1-Score	0.91	0.88	0.82	0.77

Table 7 summarizes the Inference Times across Models. The proposed approach shows that it has a speed of 35 frames per second and can definitely support real-time applications. Although Q Learning (Yu et al., 2024) also runs fast, methods Park et al. (2024) and Yang et al. (2024) show bigger processing delays, hence probably not as suitable for real-time use.

Table 7: Inference Speed Comparison of Competing Models for Real-Time Applicability.

Metric	Proposed Method	Q Learning (Yu et al., 2024)	Federated Learning (Park et al., 2024)	4DPM (Yang et al., 2024)
Frames per Second	35	34	28	24

With its properties of high accuracy, robustness, and real-time processing, it is one of the most potential candidates for practical implementations in forgery detection systems, where superior performance over state-of-the-art methods has been attained in the process. Now, we will go over an iterative, practical use case of the proposed model to help readers better understand the overall process across different scenarios.

Conclusions

The proposed framework for generalizable image and video forgery detection substantially outperforms the state-of-the-art methods on both accuracy and robustness across diverse manipulation types. The model gives very good results on the CASIA v2 image dataset and the Columbia Uncompressed Image Splicing Detection dataset with an accuracy of 96.7% and 95.8%, respectively, integrating DenseNetbased CNNs for the deep spatial feature extraction, Haralick texture features to capture handcrafted texture anomalies, and adversarially fine-tuned EfficientNet-B3. This work ensures a balanced approach by introducing an attention-based fusion mechanism between deep learning and handcrafted features that yields an F1-score of 0.95 across both datasets, outperforming methods Yu et al. (2024), Park et al. (2024), and Yang et al. (2024) by ensuring that the model marks subtle forgeries consistently with 96% recall. The robustness of the model to adversarial examples generated via GAN reinforces its practical applicability for sophisticated forgery detection, as reported at a 15-20% higher success rate against adversarial attacks compared to baseline methods. An improved spatial-temporal framework with 3D ResNet combined with LSTM to capture the long-term dependencies and TCNs for capturing short-term consistency, particularly, turned out to be successful in handling video manipulations like deepfakes. It obtained an accuracy of 94.5% on FaceForensics++ and 92.3% on DFDC, outperforming state-of-the-art works with the highest recall in deepfake detection, 95%, due to its robust dual attention mechanism, which correctly highlights the frames and regions that have been manipulated. In addition, the inference times at 30 frames per second confirm that this work is suitable for real-time applications. Fundamentally, the incorporation of multi-modal feature extraction with attention-based learning presents a holistic approach toward ever-increasing complexities in image and video forgeries.

Future Scopes

Though the proposed framework yields state-of-the-art results on several datasets, there still exists room for improvement in generalization and scalability in future works. One promising direction is to extend model capabilities using domain adaptation techniques for such domains as satellite imagery, medical imaging, and biometric authentication, where forged content may present new challenges. The reinforcement of adversarial training by more sophisticated methods of generating adversarial examples, such as using other advanced GAN architectures or other learning-based approaches for constructing adversarial examples, is another very promising direction to further develop this approach for hardening the model against adversarial attacks. Second, further extension to forgery detection in highly compressed or low-resolution images and videos where the tampering artifacts are not so visible could be done with the proposal of a novel feature enhancement technique or

generative models upsampling and refining input content to improve forgery detection. Furthermore, extension in edge-computing platforms and cloud-based real-time systems can greatly expand the scope of its applications, while it enables large-scale deployment in the area of real-time media authentication in news broadcasting, social media platforms, and government communication channels. Reasonably lightweight architectures for real-time processing while not losing accuracy will be a cardinal consideration in these applications. The evolution of deep learning techniques such as transformers might be adopted for spatial-temporal modeling in videos to capture even more subtle manipulations and further optimize this tradeoff between accuracy and computational efficiency levels. This framework lays the path to a safer digital world which, in its way, would trace and reduce the potential harm brought by image and video forgeries on a global scale during the process.

Funding

The author(s) received no specific funding for this work

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Alnaim, N. M., Almutairi, Z. M., Alsuwat, M. S., Alalawi, H. H., Alshobaili, A., & Alenezi, F. S. (2023). DFFMD: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms. *IEEE Access, 11*, 16711–16722. https://doi.org/10.1109/ACCESS.2023.3246661

Chaiwongyen, S., Duangpummet, S., Karnjana, J., Kongprawechnon, W., & Unoki, M. (2024). Potential of speech-pathological features for deepfake speech detection. *IEEE Access, 12*, 121958–121970. https://doi.org/10.1109/ACCESS.2024.3447582

Chen, H., Lin, Y., Li, B., & Tan, S. (2023). Learning features of intra-consistency and inter-diversity: Keys toward generalizable deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(3), 1468–1480. https://doi.org/10.1109/TCSVT.2022.3209336

Dong, J., Wang, Y., Lai, J., & Xie, X. (2023). Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security, 18*, 2596–2608. https://doi.org/10.1109/TIFS.2023.3266702

Huang, Y., Juefei-Xu, F., Guo, Q., Liu, Y., & Pu, G. (2023). Dodging deepfake detection via implicit spatial-domain notch filtering. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(8), 6949–6962. https://doi.org/10.1109/TCSVT.2023.3325427

Liao, X., Wang, Y., Wang, T., Hu, J., & Wu, X. (2023). FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(12), 7236–7251. https://doi.org/10.1109/TCSVT.2023.3278310

Mehra, Agarwal, A., Vatsa, M., & Singh, R. (2022). Motion magnified 3-D residual-in-dense network for deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 5*(1), 39–52. https://doi.org/10.1109/TBIOM.2022.3201887

Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access, 11*, 144497–144529. https://doi.org/10.1109/ACCESS.2023.3344653

Park, J., Park, L. H., Ahn, H. E., & Kwon, T. (2024). Coexistence of deepfake defenses: Addressing the poisoning challenge. *IEEE Access, 12*, 11674–11687. https://doi.org/10.1109/ACCESS.2024.3353785

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., ... & Vimal, V. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access*, *11*, 143296-143323. https://doi.org/10.1109/ACCESS.2023.3342107

Qiao, T., Xie, S., Chen, Y., Retraint, F., Shi, R., & Luo, X. (2024). Deepfake detection fighting against noisy label attack. *IEEE Transactions on Multimedia, 26*, 9047–9059. https://doi.org/10.1109/TMM.2024.3385286

Ramadhani, K. N., Munir, R., & Utama, N. P. (2024). Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention. *IEEE Access, 12*, 8932–8939. https://doi.org/10.1109/ACCESS.2024.3352890

Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio deepfake approaches. *IEEE Access, 11*, 132652–132682. https://doi.org/10.1109/ACCESS.2023.3333866

Tan, Z., Yang, Z., Miao, C., & Guo, G. (2022). Transformer-based feature compensation and aggregation for deepfake detection. *IEEE Signal Processing Letters, 29*, 2183–2187. https://doi.org/10.1109/LSP.2022.3214768

Waqas, N., Safie, S. I., Kadir, K. A., Khan, S., & Kaka Khel, M. H. (2022). Deepfake image synthesis for data augmentation. *IEEE Access, 10*, 80847–80857. https://doi.org/10.1109/ACCESS.2022.3193668

Wang, R., Ye, D., Tang, L., Zhang, Y., & Deng, J. (2024). AVT^2-DWF: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies. *IEEE Signal Processing Letters, 31*, 1960–1964. https://doi.org/10.1109/LSP.2024.3433596

Xie, Y., Cheng, H., Wang, Y., & Ye, L. (2024). Domain generalization via aggregation and separation for audio deepfake detection. *IEEE Transactions on Information Forensics and Security, 19*, 344–358. https://doi.org/10.1109/TIFS.2023.3324724

Xu, Y., Terhörst, P., Pedersen, M., & Raja, K. (2024). Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society, 5*(1), 93–106. https://doi.org/10.1109/TTS.2024.3365421

Yang, R., Deng, Z., Zhang, Y., Luo, X., & Lan, R. (2024). 4DPM: Deepfake detection with a denoising diffusion probabilistic mask. *IEEE Signal Processing Letters, 31*, 914–918. https://doi.org/10.1109/LSP.2024.3378127

Yang, Z., Liang, J., Xu, Y., Zhang, X.-Y., & He, R. (2023). Masked relation learning for deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 1696–1708. https://doi.org/10.1109/TIFS.2023.3249566

Yu, Y., Liu, X., Ni, R., Yang, S., Zhao, Y., & Kot, A. C. (2024). PVASS-MDD: Predictive visual-audio alignment self-supervision for multimodal deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(8), 6926–6936. https://doi.org/10.1109/TCSVT.2023.3309899

Yuan, C., Guo, Q., Zhou, Z., Fu, Z., & Xia, Z. (2023). Deepfake fingerprint detection model intellectual property protection via ridge texture enhancement. *IEEE Signal Processing Letters, 30*, 843–847. https://doi.org/10.1109/LSP.2023.3293471

Yuan, Y., Fu, X., Wang, G., Li, Q., & Li, X. (2022). Forgery-domain-supervised deepfake detection with non-negative constraint. *IEEE Signal Processing Letters, 29*, 2512–2516. https://doi.org/10.1109/LSP.2022.3193590

Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., & Tang, J. (2023). ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 1335–1348. https://doi.org/10.1109/TIFS.2023.3239223

Zhou, X., Han, H., Shan, S., & Chen, X. (2024). Fine-grained open-set deepfake detection via unsupervised domain adaptation. *IEEE Transactions on Information Forensics and Security, 19*, 7536–7547. https://doi.org/10.1109/TIFS.2024.3435440

Practical Use Case Scenario Analysis

As a proof of concept for the proposed generalizable image and video forgery detection framework, a case study pertinent to society is presented in a scenario where manipulated media-like deepfake videos and forged images are being used in spreading misinformation within an election environment. In this scenario, identification or detection of such forgeries in images and videos becomes crucial to retain trust in digital content and reduce the impact of malicious manipulation. In the case above, splicing and copy-move techniques were used to forge images while videos utilized deepfake techniques, which manipulate the appearance of political figures. The following sections show some outputs from the proposed detection models, which are: for images, DenseNet-based CNN, Haralick texture feature extraction, and adversarial fine-tuning; for video forgeries, 3D ResNet, LSTM, TCNs, and attention mechanisms are proposed. For image-based forgery detection tasks, Img001 and Img002 are splicing images sourced from the CASIA v2 dataset where parts of one image are inserted into another with an appreciable amount of subtlety in inconsistencies of texture and spatial features that arise. Img003 is a copy-move forgery, which means a portion of the same image has been copied and then pasted elsewhere within this very image, which changes the captured Haralick features of the texture pattern. Img004 and Img005 represent genuine images taken from the same dataset with no alteration, serving as the baseline for determining forged versus unchanged content. These authentic images ensure there is no generation of false positives from this model, which becomes its discriminative capability. Vid001 and Vid002 are deepfake videos acquired from the FaceForensics++ dataset for video forgery detection tasks. Deepfake techniques replace the original faces of individuals with artificially created faces. These videos contain slight visual and temporal artifacts, especially regarding facial motions, which the proposed method successfully detects. As a control example, Vid003 and Vid004 are original videos in their unaltered state taken from the same dataset. Other examples of manipulated video are Vid005, generated using the DFDC dataset, where both frame level and scene level manipulations have been made. It goes without saying that the detection of such manipulation is of immense importance for maintaining the credibility of video content in high-impact social scenarios, such as media reporting or political events. The first step treats the RGB images of 224 × 224 dimensions with the DenseNet-based CNN for deep spatial feature extraction, while the images in grayscale and their samples are used for the extraction of Haralick texture features. Then, it performs adversarial attacks on a pre-trained EfficientNet-B3 to train it more robustly. The extracted features from deep learning and handcrafted techniques are later combined by an attention mechanism to decide whether forgery has occurred or not during the process. Feature values extracted for three forged and two authentic images with corresponding forgery detection outputs are given by the following tables.

The table presents how deep features (DenseNet) combine with the texture feature of Haralick in making the final classification. In the case of three forged images, Img001, Img002, and Img003, according to the fine-tuning method of EfficientNet-B3, manipulation detection is most probable. Hence, they are confirmed as forged. In the case of the two other images, Img004 and Img005, which are authentic, the model assigns a low probability of forgery and classes them correctly. For video forgery detection, 3D ResNet for spatial analysis between successive frames, long-term temporal dependencies using LSTM, and short-term temporal consistency of TCNs were used in the process of extracting spatial-temporal features. The dual attention mechanism assigns importance to the most informative key spatial regions and temporal steps. Feature values and final classification results of video frames from a set of manipulated and authentic videos are used to assess the capability of detecting deepfake manipulations

Table 8: Generalizable Image Forgery Detection Output.

Image ID	DenseNet Feature Map (1024x7x7)	Haralick Texture Features (13-D)	EfficientNet-B3 Probability (Forgery)	Final Classification (Forgery/Authentic)
Img001	[0.23, 0.12, ..., 0.85]	[0.91, 0.12, ..., 0.65]	0.92	Forgery
Img002	[0.11, 0.34, ..., 0.67]	[0.81, 0.18, ..., 0.45]	0.88	Forgery
Img003	[0.45, 0.76, ..., 0.34]	[0.22, 0.95, ..., 0.12]	0.94	Forgery
Img004	[0.12, 0.18, ..., 0.11]	[0.01, 0.02, ..., 0.07]	0.08	Authentic
Img005	[0.14, 0.15, ..., 0.13]	[0.09, 0.05, ..., 0.04]	0.06	Authentic

The spatial-temporal features extracted for each video in the above table are through 3D ResNet, LSTM, and TCN. For manipulated videos, the attention mechanism gives more weights to both spatial and temporal features for accurate forgery classification as indicated by the Vid001, Vid002, and Vid005. To correctly classify the videos Vid003 and Vid004 as original videos, lower attention weights can be obtained. Finally, detection results are combined to assess the overall performance of the image and video forgery detection framework. In this regard, classification accuracy, precision, recall, and F1-score on test sets of both image and video datasets are enlisted in the following table. The performance of the proposed model is compared to methods Yu et al. (2024), Park et al. (2024), and Yang et al. (2024), showing the superiority of the proposed framework in detecting forgeries.

The table illustrates the very strong performance of the proposed method on both image and video datasets. In particular, the A, P, R, and F1 for images and videos are higher than those of methods Yu et al. (2024), Park et al. (2024), and Yang et al. (2024), notably improving the recall for detecting video forging processes. We provide the proof of effectiveness by combining deep spatial and handcrafted features for image forging, and the modeling of spatial-temporal features for video manipulations. It is the attention mechanisms, adversarial fine-tuning, and multi-modal feature fusion that all together create the outstanding results wowed by the proposed model process.

The last table gives the summary of the inference time both at image and video processing. These competitive times are 210 milliseconds per image, and real time at 30 fps for videos in process. The faster inference time, coupled with high accuracy, will make the proposed model more suitable for real-time and large-scale forgery detection applications.

Table 9: Enhanced Spatial-Temporal Forgery Detection Output in Videos.

Video ID	3D ResNet Spatial Features (512-D)	LSTM Temporal Features (512-D)	TCN Short-Term Temporal Features (128-D)	Attention Weights (Spatial/Temporal)	Final Classification (Forgery/Authentic)
Vid001	[0.23, 0.19, ..., 0.54]	[0.67, 0.12, ..., 0.88]	[0.18, 0.24, ..., 0.57]	[0.82 / 0.91]	Forgery
Vid002	[0.34, 0.22, ..., 0.62]	[0.45, 0.32, ..., 0.76]	[0.22, 0.16, ..., 0.64]	[0.78 / 0.87]	Forgery
Vid003	[0.15, 0.45, ..., 0.35]	[0.21, 0.14, ..., 0.28]	[0.17, 0.18, ..., 0.45]	[0.34 / 0.42]	Authentic
Vid004	[0.11, 0.09, ..., 0.14]	[0.09, 0.12, ..., 0.13]	[0.08, 0.07, ..., 0.16]	[0.12 / 0.14]	Authentic
Vid005	[0.2, 0.21, ..., 0.65]	[0.59, 0.41, ..., 0.77]	[0.25, 0.29, ..., 0.68]	[0.79 / 0.84]	Forgery

Table 10: Overall Performance Metrics for Image and Video Forgery Detection.

Metric	Proposed Method (Images)	Q Learning (Yu et al., 2024) (Images)	Federated Learning (Park et al., 2024) (Images)	4DPM (Yang et al., 2024) (Images)	Proposed Method (Videos)	Q Learning (Yu et al., 2024) (Videos)	Federated Learning (Park et al., 2024) (Videos)	4DPM (Yang et al., 2024) (Videos)
Accuracy (%)	96.7	94.1	90.8	88.3	94.5	91.3	88.5	85.0
Precision	0.95	0.93	0.91	0.89	0.94	0.92	0.88	0.84
Recall	0.96	0.89	0.87	0.81	0.95	0.91	0.86	0.80
F1-Score	0.95	0.91	0.89	0.84	0.94	0.91	0.87	0.82

Table 11: Inference Time for Image and Video Processing.

Metric	Proposed Method (Images)	Q Learning (Yu et al., 2024) (Images)	Federated Learning (Park et al., 2024) (Images)	4DPM (Yang et al., 2024) (Images)	Proposed Method (Videos)	Q Learning (Yu et al., 2024) (Videos)	Federated Learning (Park et al., 2024) (Videos)	4DPM (Yang et al., 2024) (Videos)
Inference Time (ms)	210	235	280	300	30 fps	34 fps	28 fps	24 fps