Lakshay Bhardwaj, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Multi-Branch Deepfake Detection Framework Using Spatial, Noise, and Frequency Domain Features with Attention Fusion

Lakshay Bhardwaj[1], Rishabh Jain[2], Kritika[3], Ritesh Kumar[4]
[1,2,3]B.Tech.(AI&DS) 2nd Yr Scholar, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi
[4]Assistant Professor, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi

Abstract- Deepfakes have emerged as a major challenge in digital media forensics because modern generative models can produce highly realistic fake facial content that is difficult to distinguish from authentic media. Their rapid growth has increased the risk of misinformation, identity misuse, and security threats in online environments, motivating the need for reliable forensic datasets and detection frameworks [1], [7], [10].
Existing deepfake detection methods often perform well only under controlled settings and show limited robustness when evaluated on unseen manipulations or datasets. Many CNN-based methods achieve high accuracy on known datasets but fail to generalize to unseen data. Prior works based on compact CNNs, frequency-aware learning, and multi-branch detection have shown promising performance, but cross-dataset generalization remains a major challenge [2]–[5], [8]. To address this issue, this work proposes a multi-branch deepfake detection framework that jointly learns from spatial appearance information, residual noise traces, and frequency-domain artifacts. The spatial branch uses a pretrained EfficientNet-B0 backbone to capture facial inconsistencies [6], the noise branch extracts forensic residual cues using SRM-based filtering inspired by image manipulation detection methods [9], and the frequency branch analyzes the log magnitude spectrum obtained through FFT transformation to reveal spectral anomalies commonly associated with forged content [3]. An attention-based fusion module combines these complementary representations and adaptively emphasizes the most discriminative branch for each sample, following the motivation of prior multi-domain and multi-branch approaches [4], [5]. The model is trained and evaluated on the FaceForensics++ dataset using frame-level samples derived from video sequences [1]. Experimental results show that the proposed framework achieves a final test accuracy of 63.75%, demonstrating that multi-domain feature fusion is effective for improving deepfake detection performance. The results further indicate that attention-guided fusion helps the classifier exploit complementary forensic evidence beyond conventional RGB-only models.

Keywords: Deepfake detection, Digital media forensics, Generative models, Fake facial content, Deep learning, Convolutional Neural Networks (CNNs).

I. INTRODUCTION

Background

Deepfakes are synthetically manipulated images or videos in which facial identity, expressions, or speech-related visual patterns are altered using deep learning techniques. With the rapid advancement of generative adversarial networks, autoencoders, and neural rendering methods, deepfakes have become increasingly realistic and easier to produce at scale. Public datasets such as FaceForensics++ and DFDC illustrate both the progress of synthesis quality and the growing need for reliable detection systems [1], [10].

At the same time, several detection frameworks have been proposed to address this challenge, ranging from compact convolutional models to multi-domain forensic approaches. Earlier works such as MesoNet demonstrated the feasibility of CNN-based facial forgery detection, while later approaches explored frequency-aware and multi-branch learning to capture more subtle manipulation traces [2]–[5]. Efficient deep learning backbones such as EfficientNet have further supported the development of accurate and computationally practical detection architectures [6].

Problem Statement

The misuse of deepfakes creates serious risks in fake news dissemination, political misinformation, social manipulation, identity fraud, and biometric security attacks. As synthesis methods improve, visible artifacts become less obvious, making manual verification unreliable and automated detection increasingly difficult. High-quality datasets such as Celeb-DF show that many earlier detectors struggle when fake content contains fewer obvious visual defects [7].

In real-world scenarios, deepfake detection systems must operate under varying levels of compression, diverse lighting conditions, changing facial poses, and continuously evolving generation methods. This makes it difficult for conventional detectors to maintain stable performance outside the specific datasets on which they are trained [7], [10].

Limitations of Existing Methods

A large number of deepfake detection systems rely mainly on CNN-based spatial feature extraction. While these methods can achieve strong performance on benchmark datasets, they often learn dataset-specific artifacts instead of manipulation-invariant forensic cues. Many CNN-based methods achieve high accuracy on known datasets but fail to generalize to unseen data. Cross-dataset studies have shown that performance can degrade significantly because of domain shift, compression variation, manipulation diversity, and differences in video quality [2], [7], [8].

Although frequency-based methods have shown promise in revealing hidden spectral inconsistencies [3], and multi-branch or cross-domain models have improved feature representation by combining complementary cues [4], [5], the generalization problem remains far from solved. In addition, forensic noise-based methods inspired by image manipulation detection have shown that residual traces can be useful, but such cues are not always sufficiently exploited in mainstream deepfake detectors [9]. These limitations indicate that relying on a single feature domain is often inadequate for robust detection.

Objective of the Work

The objective of this work is to build a robust deepfake detection model that does not depend only on visual appearance cues. Instead, it integrates spatial, noise-based, and frequency-domain information in a unified architecture to improve robustness and generalization. The proposed model is designed to exploit complementary forensic traces so that it can better distinguish authentic media from manipulated content even when visible artifacts are subtle.

To achieve this, the proposed framework combines an EfficientNet-based spatial branch for semantic visual analysis [6], a residual noise branch motivated by forensic filtering strategies [9], and a frequency-analysis branch inspired by frequency-aware deepfake detection methods [3]. These branches are fused within a multi-branch architecture motivated by prior work on cross-domain and two-stream deepfake detection [4], [5], with the aim of improving performance on benchmark datasets such as FaceForensics++ and enhancing robustness beyond conventional CNN-only approaches [1], [8].

II. LITERATURE REVIEW

CNN-Based Methods

Early deepfake detection research was dominated by convolutional neural network (CNN)-based methods, mainly because CNNs had already shown strong performance in image classification, face analysis, and digital image forensics. In the context of deepfake detection, the central idea behind CNN-based approaches is that manipulated media often contains subtle spatial inconsistencies, blending defects, unnatural textures, or boundary artifacts that can be learned directly from pixel-level inputs. Instead of manually engineering forensic descriptors, CNNs automatically learn hierarchical features that distinguish authentic facial content from forged content. This learning-based paradigm made CNNs the most natural starting point for deepfake detection systems [1], [2].

One of the earliest and most influential CNN-based methods for facial forgery detection is MesoNet, proposed by Afchar et al. [2]. MesoNet was designed

as a compact and computationally efficient network that focuses on mesoscopic image properties rather than extremely fine-grained pixel noise or very high-level semantic structure. The authors argued that deepfake artifacts often lie in an intermediate representational space: they are not always obvious enough to be detected through handcrafted visual inspection, yet they do not necessarily require very deep semantic reasoning either. Based on this insight, MesoNet introduced lightweight CNN architectures capable of identifying manipulated facial videos efficiently. The significance of MesoNet lies not only in its performance but also in its practicality. It demonstrated that effective deepfake detection did not always require extremely large models and that compact architectures could still capture discriminative forgery traces when trained properly [2].

As the field matured, larger and deeper CNNs began to outperform compact architectures on benchmark datasets. A major milestone was the introduction of FaceForensics++ by Rössler et al. [1], which provided one of the most widely used benchmarks for manipulated facial imagery. FaceForensics++ contained a large-scale dataset of original and manipulated videos generated using multiple face manipulation methods, including DeepFakes, Face2Face, FaceSwap, and NeuralTextures. More importantly, it established a standardized evaluation framework for deepfake detection. The authors showed that CNN-based detectors, especially deeper architectures such as Xception, could achieve strong performance on known manipulations and under controlled benchmarking conditions [1]. This benchmark played a crucial role in accelerating the field because it allowed researchers to compare different CNN architectures on a common dataset rather than relying on limited or custom data.

However, the strong performance of CNN-based detectors on FaceForensics++ also highlighted an important limitation. Many of these models performed well when the training and testing data came from the same distribution, but their performance dropped when evaluated on media generated by different synthesis techniques or from different datasets. In other words, they often learned dataset-specific artifacts instead of universal manipulation cues. This concern became more visible with the release of Celeb-DF, a more challenging dataset proposed by Li et al. [7]. Celeb-DF was created to address the overly obvious artifacts present in earlier datasets and to better reflect the visual quality of modern deepfake generation. The authors showed that several existing CNN-based detection methods, which reported strong performance on prior benchmarks, struggled significantly on Celeb-DF. This demonstrated that good performance on one dataset did not necessarily translate into robustness in real-world scenarios [7].

The issue of generalization became even more pronounced with the introduction of the DeepFake Detection Challenge (DFDC) dataset by Dolhansky et al. [10]. DFDC is substantially larger and more diverse than many earlier benchmarks and includes wide variation in subjects, recording conditions, lighting, pose, and generation quality. The dataset emphasized the fact that deepfake detection is not merely a binary classification task under laboratory conditions but a real-world forensic challenge involving substantial domain variability. CNN-based methods trained on narrow or homogeneous datasets often fail to cope with this diversity. As a result, the literature increasingly recognized that architecture design alone is not sufficient; the diversity and realism of the training data strongly influence the generalization of CNN-based detectors [10].

Another important development in CNN-based detection is the use of efficient backbone networks such as EfficientNet. Although EfficientNet was not proposed specifically for deepfake detection, Tan and Le introduced a family of models that scale network depth, width, and resolution in a principled manner, achieving an excellent trade-off between accuracy and efficiency [6]. This is highly relevant to deepfake detection because practical forensic systems must often balance detection accuracy with computational cost. EfficientNet-based backbones have become attractive for deepfake detection frameworks because they can extract strong visual features while remaining lighter than many

conventional very-deep CNNs. Their efficiency is especially valuable in frame-based video analysis, large-scale dataset training, and deployment-oriented applications [6].

Despite these advances, a recurring weakness of CNN-based methods is their tendency to exploit superficial visual artifacts. Compression noise, resizing patterns, color mismatch, and low-resolution face warping may unintentionally become shortcuts that the network uses for classification. When these cues disappear in higher-quality deepfakes, the detector may fail. Nadimpalli and Rattani explicitly addressed this concern in their work on cross-dataset generalization, showing that detectors with high in-dataset performance can degrade substantially in cross-dataset evaluation due to domain shift [8]. Their findings reinforce a critical conclusion from the CNN-based literature: although CNNs are highly effective at learning patterns from large datasets, they do not automatically learn robust forensic representations unless the training strategy and data diversity explicitly encourage generalization [8].

Overall, CNN-based methods form the foundation of deepfake detection research. They established the feasibility of automatic forgery detection, enabled large-scale benchmarking, and provided strong baseline performance. However, the limitations revealed by Celeb-DF, DFDC, and cross-dataset studies indicate that spatial CNN features alone are often insufficient for robust real-world detection. These shortcomings motivated later work to move beyond pure RGB-based analysis and explore additional signal domains such as frequency and residual noise [1], [2], [7], [8], [10].

Frequency-Based Methods

Frequency-based deepfake detection emerged from the observation that synthetic image generation often leaves artifacts in the transformed domain that may not be obvious in the original RGB image. While a deepfake may appear visually realistic to a human observer, its generation pipeline can disturb the natural spectral distribution of real images. Interpolation, upsampling, blending, and generative synthesis can all introduce anomalies in high-frequency or band-specific components. Frequency-domain analysis therefore offers a complementary perspective to spatial CNN analysis by focusing on how energy is distributed across image frequencies rather than how the image appears directly in pixel space [3].

A seminal contribution in this direction is F3-Net, proposed by Qian et al. [3]. The central idea of F3-Net is that face forgeries contain frequency-aware clues that standard RGB-domain networks may overlook. To exploit this, the authors designed a network that captures two complementary types of frequency information: frequency-aware decomposed image components and local frequency statistics. Rather than simply applying a transform and feeding the result to a classifier, F3-Net carefully models the transformed-domain characteristics of forged content and integrates them into a collaborative learning framework. This design reflects an important shift in the literature from treating frequency representations as auxiliary handcrafted features to learning them in an end-to-end deep architecture [3].

The importance of F3-Net lies in both its conceptual and practical contributions. Conceptually, it demonstrated that the frequency domain is not just a pre-processing space but a meaningful forensic representation where manipulated content may be more separable from authentic content. Practically, it showed that explicitly modeled frequency clues can improve detection performance, especially when visual artifacts in the RGB domain are weak. This insight became increasingly valuable as deepfake synthesis quality improved. As datasets evolved from earlier, artifact-heavy manipulations to cleaner and more realistic forgeries, the need for transformed-domain analysis became more apparent [3], [7].

The motivation for frequency analysis is also connected to the weaknesses of purely spatial CNNs. Spatial networks may focus on semantic content, facial identity, or dataset-specific appearance cues rather than manipulation traces. In contrast, frequency-based features can be more sensitive to unnatural synthesis operations such as repetitive patterns from generative models, inconsistencies in

edge transitions, and interpolation-induced distortions. These patterns may remain detectable even when the fake face looks natural to the eye. For this reason, frequency-domain methods are often viewed as a way to reveal "hidden" forensic evidence that complements visible spatial features [3].

The relevance of frequency analysis is also reflected in other multi-domain models. For example, the two-branch recurrent network proposed by Masi et al. [4] incorporates a branch that emphasizes artifact amplification using frequency-related filtering while suppressing dominant facial content. Although the model is framed as a two-branch video-level detector rather than a pure frequency-domain architecture, its design supports the broader argument that frequency-enhanced representations help isolate manipulative traces more effectively than RGB inputs alone. By combining this with recurrent temporal modeling, the method captures both artifact patterns and temporal consistency across video frames [4].

Similarly, MD-CSDNetwork by Agarwal et al. [5] integrates RGB and frequency-domain representations using cross-stitch units. In this framework, one branch processes spatial information while another processes frequency information, and the network learns how much knowledge should be shared between them. This design illustrates an important development in the literature: frequency analysis is increasingly not treated as a separate branch standing in opposition to CNN features but as a cooperative domain that enhances CNN-based learning. Instead of choosing between spatial and transformed-domain analysis, newer methods attempt to combine both in a way that preserves the strengths of each [5].

The importance of frequency-domain methods also becomes clearer when examined against challenging datasets. In datasets like FaceForensics++, some manipulations still contain visible artifacts that strong spatial CNNs can exploit [1]. But in datasets such as Celeb-DF, where synthesis quality is improved and obvious defects are reduced, subtle transformed-domain irregularities may become more informative than direct pixel appearance [7]. This suggests that frequency-based analysis is especially useful for high-quality forgeries, where realism in the spatial domain does not necessarily imply realism in the spectral domain.

Another advantage of frequency-based features is their potential relevance to robustness against compression and post-processing. Social media platforms often compress videos heavily, which can degrade spatial artifacts and reduce the reliability of RGB-only cues. Although compression also affects the frequency domain, certain spectral inconsistencies may remain relatively informative after degradation. This makes frequency-aware detectors appealing for practical scenarios where manipulated videos undergo multiple stages of re-encoding, resizing, or transmission [1], [3].

Nevertheless, frequency-based methods are not without challenges. First, transformed-domain features can also be affected by dataset bias, compression characteristics, and acquisition pipelines. A model may learn spectral patterns associated with a dataset rather than a manipulation process. Second, frequency representations alone may not capture enough semantic information to discriminate authentic and fake faces in all cases. This is why most successful recent methods do not rely exclusively on frequency features but instead combine them with spatial or temporal information. Third, the best choice of transform—FFT, DCT, local frequency statistics, or band decomposition—remains an active design consideration, and different transforms may emphasize different kinds of forensic traces [3], [5].

In summary, frequency-based deepfake detection has become a major branch of the literature because it addresses an important limitation of RGB-only CNN methods. By analyzing spectral irregularities introduced during face synthesis and manipulation, these approaches provide complementary evidence that is often more robust when deepfakes become visually convincing. The work of Qian et al. [3], along with related multi-domain approaches [4], [5], established frequency-aware learning as a key component in modern deepfake forensics.

Noise-Based Methods

Noise-based deepfake detection methods originate from the broader field of image forensics, where forensic analysts have long used residual noise, sensor patterns, and local inconsistencies to identify tampering. The underlying principle is that manipulated regions often disturb the natural statistical properties of an image. Even when a forged image appears visually coherent, the generation or editing process can alter local noise distributions, disrupt sensor-related traces, or create inconsistent residual patterns. These signals are often too subtle to be interpreted directly in the RGB domain, which is why residual extraction and noise enhancement techniques are commonly applied before learning-based analysis [9].

A foundational reference for this direction is the work of Zhou et al. on Learning Rich Features for Image Manipulation Detection [9]. Although the paper addresses image manipulation detection more broadly rather than deepfakes specifically, it is highly relevant because it introduces the use of SRM-inspired filtering as a way to capture local noise inconsistencies. The authors proposed a two-stream approach in which one stream processes RGB appearance while the other processes residual noise maps generated using selected SRM filters. The key insight is that manipulation artifacts may be easier to detect after suppressing semantic image content and emphasizing residual traces. This idea has had a strong influence on later deepfake detection methods because deepfake generation, like other forms of image tampering, can disturb local statistical regularities even when overall facial appearance looks realistic [9].

SRM, or Spatial Rich Model filtering, was originally developed in steganalysis and forensic analysis to capture subtle pixel-level dependencies. In deepfake detection, SRM-inspired preprocessing helps expose discrepancies caused by face blending, interpolation, synthesis, and compression. These artifacts may not form easily recognizable objects or textures, but they can appear as unnatural high-frequency perturbations, smoothing patterns, or local inconsistencies in residual space. By feeding these residual maps to a CNN, the detector learns to focus less on face identity and more on the forensic fingerprints of manipulation [9].

Noise-based methods are especially important because of the weaknesses of pure semantic learning. A spatial CNN trained directly on RGB faces may implicitly learn identity features, skin tones, backgrounds, or dataset-specific facial framing. Such shortcuts are problematic because they do not reflect manipulation evidence. In contrast, residual-domain processing attempts to strip away some of this semantic dominance and redirect the network toward low-level forensic cues. This can improve robustness, particularly when fake and real samples share similar visual appearance but differ in local pixel statistics [9].

The relevance of noise-based analysis is also consistent with findings from deepfake benchmarks. FaceForensics++ demonstrated that deep learning methods could outperform human observers, but it also showed performance sensitivity to compression and manipulation type [1]. This sensitivity suggests that models need access to subtle forensic signals beyond visible appearance. In higher-quality datasets such as Celeb-DF, where color mismatch and blending boundaries are reduced, residual inconsistencies may provide an alternative source of evidence when obvious spatial artifacts are absent [7].

Noise-based approaches also relate closely to the generalization problem. Nadimpalli and Rattani showed that deepfake detectors often struggle across datasets because they overfit dataset-specific properties [8]. Residual-based features may help mitigate this issue to some extent by emphasizing manipulation-induced traces instead of semantic content. However, this is not guaranteed. Residual statistics themselves can still depend on compression pipeline, camera characteristics, and dataset acquisition conditions. Therefore, while noise-based methods are promising, their success still depends on training diversity and careful evaluation [8].

In practice, noise-based detection is rarely used in isolation. Most modern methods combine residual analysis with spatial or frequency features. This is

because residual features are highly informative about local inconsistencies but may not capture enough global structural context. For example, a residual map may reveal an unnatural boundary but may not fully indicate whether the overall facial geometry is plausible. Combining noise features with spatial and frequency cues therefore provides a more complete forensic picture. This design philosophy is reflected in recent multi-branch architectures and is one of the core motivations behind combining SRM-based branches with CNN backbones and spectral modules [4], [5], [9].

Overall, noise-based methods occupy an important position in the literature because they bridge classical image forensics and modern deep learning. They provide a principled way to capture subtle manipulation traces that RGB-based models may miss. The use of SRM-inspired filtering, in particular, has become a widely accepted strategy for exposing forensic residual patterns and has directly influenced multi-branch deepfake detection frameworks [9].

Multi-Modal Approaches

As the limitations of single-domain detection became more evident, the literature increasingly shifted toward multi-modal or multi-domain approaches. The core motivation behind these methods is that deepfakes leave evidence in more than one representational space. A manipulated face may appear suspicious in RGB texture, exhibit abnormal residual noise, and contain unnatural frequency patterns simultaneously. A detector that relies on only one of these cues risks missing important evidence or overfitting to dataset-specific artifacts. Multi-modal approaches attempt to address this by combining complementary features so that weaknesses in one domain can be compensated by strengths in another [3]–[5].

One influential example is the Two-branch Recurrent Network for Isolating Deepfakes in Videos proposed by Masi et al. [4]. This model combines color-domain information with artifact-amplified features and then applies recurrent temporal modeling to make video-level predictions. The architecture is particularly important because it recognizes that deepfake detection in videos is not only a spatial problem but also a temporal one. Manipulation inconsistencies may appear across consecutive frames, and modeling these temporal relationships can improve robustness. By using a dual-branch design, the method suppresses dominant facial content while highlighting artifacts more relevant to manipulation. This reflects a broader trend in the literature: effective deepfake detection often requires selective enhancement of forensic evidence rather than simple end-to-end classification from raw RGB frames [4].

Another important multi-domain model is MD-CSDNetwork by Agarwal et al. [5]. This network combines RGB and frequency-domain inputs using cross-stitch units that learn how much information should be shared between branches. The key innovation here is adaptive feature sharing. Instead of manually concatenating features or assigning fixed fusion weights, the network learns interactions between domains during training. This allows the model to exploit both domain-specific and shared representations. Such adaptive fusion is particularly appealing in deepfake detection because the relative usefulness of spatial and frequency cues may vary from sample to sample [5].

F3-Net can also be viewed as part of this multi-modal evolution because it combines different kinds of frequency-aware clues within a unified architecture [3]. Although its focus is frequency-domain learning, its branch design illustrates a general methodological trend: complementary representations should not be treated independently, but fused through learnable interactions. This idea has influenced later attention-based and branch-fusion models, where the system learns the contribution of each domain dynamically rather than statically [3].

The rise of multi-modal methods is closely connected to the generalization problem highlighted by Nadimpalli and Rattani [8]. If a detector relies only on one type of cue, such as visual artifacts or spectral anomalies, it may overfit to that specific cue and fail when it is absent in a new dataset. By contrast, multi-domain systems provide redundancy. If spatial artifacts are weak, frequency

or residual cues may still reveal manipulation. If compression suppresses high-frequency details, spatial inconsistencies may still remain useful. This redundancy is one of the main reasons multi-modal detectors are seen as more promising for robust deepfake forensics [8].

Benchmark datasets also indirectly support the move toward multi-modal learning. FaceForensics++ enabled strong CNN-based baselines but did not solve cross-manipulation robustness [1]. Celeb-DF showed that visually realistic deepfakes can defeat many earlier detectors [7]. DFDC emphasized large-scale diversity and real-world variation [10]. Together, these datasets revealed that no single cue is sufficient across all settings. This has encouraged researchers to combine multiple evidence sources, including RGB appearance, residual noise, spectral statistics, and temporal consistency [1], [7], [10].

From an architectural standpoint, multi-modal methods differ mainly in how they extract and fuse features. Some methods use parallel branches followed by concatenation; others use cross-attention, cross-stitch units, bilinear pooling, or recurrent fusion. The common objective, however, is to preserve domain-specific strengths while allowing interaction across branches. In deepfake detection, this is especially important because manipulation traces are often subtle, heterogeneous, and context-dependent. A fixed rule-based fusion may not work equally well for all samples, whereas adaptive fusion can learn to emphasize the most informative branch in each case [4], [5].

In summary, multi-modal approaches represent a natural and necessary progression in deepfake detection research. They build on the strengths of CNN-based, frequency-based, and noise-based methods while addressing many of their individual limitations. The literature suggests that robust detection is more likely to emerge from integrating complementary domains than from pursuing a single feature type in isolation. This directly motivates the present work, which extends prior multi-branch ideas by explicitly combining spatial, noise, and frequency-domain representations through an attention-based fusion mechanism [3]–[5].

III. PROPOSED METHODOLOGY

Overall System Pipeline

The proposed deepfake detection framework is designed to learn complementary forensic cues from three different domains: spatial appearance, residual noise, and frequency information. The complete pipeline begins with a labeled video dataset containing both authentic and manipulated face videos. Publicly available benchmark datasets such as FaceForensics++, Celeb-DF, and DFDC have demonstrated that deepfake detection systems must operate under diverse compression levels, visual qualities, and forgery types [1], [7], [10]. Motivated by these challenges, the proposed system converts videos into representative frame samples and performs classification at the frame level.

In the first stage, frames are extracted from each video and organized into a structured image dataset. These extracted frames are then grouped according to their source videos and divided into training and testing sets using a video-level splitting strategy. After splitting, data augmentation is applied to the training set in order to simulate realistic variations such as cropping, compression, blur, and illumination change. Each preprocessed image is then passed through three parallel feature extraction branches. The first branch captures spatial artifacts using a pretrained EfficientNet-B0 backbone [6], the second branch highlights residual forensic traces through SRM-based filtering inspired by image forensics literature [9], and the third branch analyzes spectral inconsistencies through Fast Fourier Transform (FFT)-based frequency representations [3].

The outputs of these three branches are combined using an attention-based fusion module that adaptively learns the relative importance of spatial,noise, and frequency cues for a given input sample. The fused representation is then passed through fully connected layers for binary classification into real and fake classes. Finally, the trained model is evaluated using standard

classification metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. This multi-branch design is motivated by prior studies showing that spatial-only models may miss important forensic artifacts, while multi-domain approaches often achieve stronger robustness and better discrimination [3]–[5], [8].

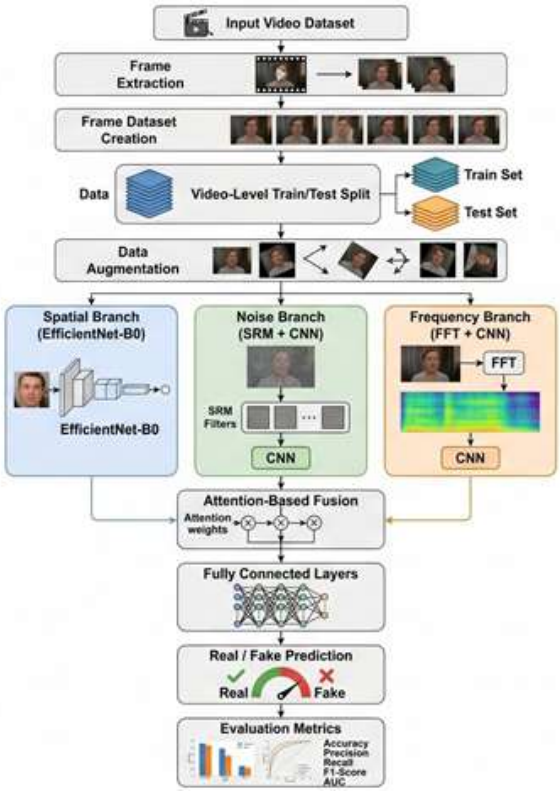

Figure 1. Overall pipeline diagram

Frame Extraction Process

Deepfake datasets are generally distributed in video form, but training directly on complete videos requires high computational resources and more complex temporal modeling. To reduce this burden while preserving useful forensic content, the proposed method converts each video into a fixed number of representative image frames. This frame-based strategy is widely used in deepfake detection because it increases the number of training samples, lowers memory consumption, and allows efficient use of image-based CNN architectures [1], [2], [4].

For each video, a fixed number of frames is extracted through uniform temporal sampling. Instead of selecting only consecutive or initial frames, the extraction process spreads sampled frames across the full duration of the video. A linspace based indexing method is used to generate evenly spaced frame positions, ensuring that the extracted set captures visual content from the beginning, middle, and end of the clip. This reduces temporal sampling bias and helps preserve manipulation artifacts that may vary across the video.

The use of frames instead of full videos offers several advantages. First, image-based learning is computationally more efficient than full spatiotemporal video modeling. Second, frame extraction allows the dataset size to be expanded considerably without collecting additional videos. Third, many deepfake artifacts such as blending errors, texture inconsistencies, abnormal boundaries, and local spectral distortions can already be observed at the frame level [1], [3], [9]. Thus, frame-wise analysis offers a practical balance between detection accuracy and implementation complexity.
To improve preprocessing reliability, the extraction pipeline incorporates several safeguards. If a video file contains unreadable or corrupted frames, those frames are skipped rather than causing the entire extraction process to fail. In addition, extraction progress is recorded in log files so that processing can resume from the last completed video in case of interruption. This resumable design is especially useful for large-scale datasets such as FaceForensics++ and DFDC, where preprocessing may involve thousands of videos [1], [10].

Dataset Splitting Strategy

After frame extraction, the dataset is divided into training and testing subsets. In the proposed work, the split is performed at the video level, not at the frame level. To avoid data leakage, splitting is performed at video level rather than frame level. This is a critical design choice because frames extracted from the same video are highly correlated in facial appearance, background, lighting, and compression characteristics. If some frames from a video appear in the training set while other frames from the same video appear in the test set, the model may benefit from memorizing video-specific patterns instead of learning generalizable manipulation cues.

Video-level splitting ensures that all frames originating from one source video remain within a single subset only. As a result, the evaluation becomes more realistic and better reflects the model's actual ability to generalize to unseen videos. This issue is particularly important in deepfake detection research, where overly optimistic performance can arise from improper dataset partitioning. Prior benchmark studies have shown that evaluation protocols strongly affect reported performance, especially when models are trained on large datasets such as FaceForensics++ and then tested on more challenging datasets such as Celeb-DF [1], [7], [8].

The proposed strategy also supports future cross-dataset experimentation. Since poor cross-dataset generalization remains one of the main limitations of existing detectors, careful within-dataset splitting is necessary before extending experiments to unseen domains [8]. By maintaining strict separation between source videos, the present work follows a sound experimental protocol and reduces the risk of inflated accuracy estimates.

Data Augmentation Techniques

Deepfake detection models often suffer from limited generalization because they may overfit to dataset-specific artifacts such as resolution, compression, or generation quality. To reduce this problem, the proposed framework applies a set of data augmentation techniques during training. These augmentations expose the model to diverse visual distortions and help it learn more robust manipulation features rather than memorizing superficial cues. This idea is also consistent with prior work emphasizing the importance of robustness and generalization in unseen settings [7], [8].

The following augmentations are used:

1. Random Resized Crop: introduces variation in scale and framing. This helps the model become less sensitive to exact face position and crop boundaries.
2. Horizontal Flip: improves left-right invariance and prevents the model from associating forgery cues with a fixed facial orientation.
3. Small Rotation: makes the network more robust to minor head tilt, pose variation, and imperfect face alignment.
4. Color Jitter: Simulates changes in illumination, contrast, brightness, and saturation. This is useful because videos collected in practical scenarios may differ significantly in lighting conditions.
5. Gaussian Blur: introduces controlled degradation and helps the model handle low-quality or motion-affected frames.
6. JPEG Compression: simulates real-world social media recompression and transmission artifacts. JPEG compression augmentation improves robustness to real-world deepfake artifacts.

Among these, JPEG compression is particularly important because many deepfake videos are shared through online platforms where repeated compression may hide or distort high-frequency artifacts. Previous benchmark datasets such as FaceForensics++ explicitly evaluate detection under different compression levels, and challenging datasets like Celeb-DF show that detectors trained only on clean artifacts often struggle on higher-quality forgeries [1], [7]. Therefore, augmentation acts as a regularization mechanism that improves the model's ability to detect manipulations under realistic degradation conditions.

Proposed Multi-Branch Model Architecture

The core contribution of this work is a multi-branch architecture that jointly models spatial, residual noise, and frequency-domain evidence. Deepfake generation methods may leave traces in more than one domain. Some manipulations introduce visible semantic inconsistencies, while others alter local noise statistics or generate spectral artifacts that are difficult to detect in RGB space alone. Recent research has shown that frequency-aware and multi-domain representations can significantly strengthen forgery detection [3]–[5]. Motivated by this, the proposed model extracts features from three complementary branches and combines them through an adaptive fusion mechanism.

Spatial Branch

The spatial branch is responsible for learning high-level visual and semantic inconsistencies directly from the RGB face image. In this work, a pretrained EfficientNet-B0 model is used as the backbone network. EfficientNet-B0 is selected because it provides an effective balance between model size, computational efficiency, and feature extraction capability through compound scaling of depth, width, and input resolution [6]. These properties make it suitable for deepfake detection systems that aim to achieve good performance without excessive computational cost.

The spatial branch focuses on capturing visible manipulation artifacts such as :

- inconsistent skin texture,
- unnatural facial boundaries,
- blending errors between source and target face regions,
- warping distortions,
- local appearance mismatch,
- abnormal shading or facial semantics.

CNN-based deepfake detectors such as MesoNet and the FaceForensics++ benchmark have demonstrated that strong spatial backbones can detect many types of face manipulation artifacts effectively [1], [2]. However, RGB-based models may still be vulnerable when manipulations become visually realistic or when testing conditions differ from the training dataset. For this reason, the spatial branch in the proposed model is not used alone, but rather as one component of a broader multi-domain framework.

Feature vectors extracted from the final convolutional representation of EfficientNet-B0 are forwarded to the fusion stage. These features represent semantic and textural evidence and serve as the primary visual stream of the network [6].

Noise Branch

The noise branch is designed to capture subtle residual traces that are not easily visible in the original RGB image. In digital image forensics, residual-based representations have proven effective for exposing traces of tampering, resampling, interpolation, and post-processing. Following this idea, the proposed method uses SRM (Spatial Rich Model) filters as a preprocessing step to suppress image content and highlight noise residuals [9].

SRM filtering transforms the input image into a residual domain in which local pixel dependencies, fine-grained inconsistencies, and manipulation artifacts become more prominent. Such residual patterns may arise from face blending, boundary correction, GAN synthesis, or compression. The use of SRM-inspired features in forensic detection has been shown to improve manipulation analysis by focusing less on semantic content and more on low-level statistical irregularities [9].

After SRM filtering, the residual maps are passed through a CNN-based subnetwork that learns higher-order residual representations. This CNN is responsible for identifying discriminative forensic patterns from the noise-enhanced input. In the context of deepfake detection, the branch helps the system detect artifacts that may remain even when the synthesized face looks perceptually convincing. Similar ideas of combining appearance-based and residual-based analysis have been shown to improve manipulation detection in prior forensic studies [4], [9].

The main role of the noise branch is therefore to complement the spatial branch. While the spatial stream focuses on what the image looks like, the noise stream focuses on how the image was formed and whether its local residual characteristics are consistent with authentic content.

Frequency Branch

The frequency branch analyzes the input image in the transformed spectral domain. Deepfake generation and post-processing operations often introduce frequency distortions that are not obvious in pixel space but become visible after transformation. Prior work such as F3-Net has shown that frequency-aware learning can reveal manipulation clues that conventional RGB-based models may miss [3].

In the proposed framework, the input image is first converted to the frequency domain using the Fast Fourier Transform (FFT). From the FFT output, the log magnitude spectrum is computed and used as the branch input. This spectrum highlights how energy is distributed across different frequency bands and helps expose abnormal periodic structures, high-frequency inconsistencies, and unnatural synthesis patterns.

The frequency branch is then implemented using a CNN module that learns discriminative representations from the spectrum image. Its goal is to distinguish the spectral properties of real frames from those of forged frames. Unlike the spatial branch, which focuses on semantic appearance, the frequency branch captures statistical regularities that reflect the generation process itself. This makes it especially valuable when fake images have visually plausible facial structure but still contain hidden frequency-domain artifacts [3].

The use of frequency information is further supported by multi-domain deepfake detection research, including two-stream and cross-stitched networks that combine RGB and transformed-domain features to improve detection robustness [3]–[5]. In the proposed model, the FFT branch serves as a complementary cue source that enhances the overall representation.

Attention-Based Fusion

After feature extraction, the outputs of the spatial, noise, and frequency branches are combined using an attention-based fusion module. The motivation behind this design is that the relative importance of each branch may vary from sample to sample. For example, some fake frames may contain strong visible blending artifacts, while others may appear visually smooth but still reveal residual or spectral inconsistencies. Therefore, assigning fixed equal importance to all three branches may not be optimal. The attention module learns adaptive weights for the three branch-level feature vectors. Let the extracted features from the spatial, noise, and frequency branches be denoted by fs, fn and ff, respectively. The fusion module estimates attention coefficients αs,αn and αf, where:

$\alpha s+\alpha n+\alpha f=1$

The fused representation is then computed as:
fusion = $\alpha s fs + \alpha n fn + \alpha f ff$
This strategy enables the network to emphasize the most informative domain for each input instance. Similar ideas of feature interaction and cross-domain integration have been shown to be effective in frequency-aware and multi-domain deepfake detection frameworks [3], [5]. In the proposed work, attention fusion improves interpretability and allows the classifier to exploit complementary evidence more effectively than a single-stream architecture.

Classification Layer

The fused representation is passed through one or more fully connected layers to produce the final binary decision. These dense layers serve two purposes: first, they integrate the multi-domain information into a compact discriminative representation; second, they map the learned features into class probabilities corresponding to real and fake.

A nonlinear activation function such as ReLU is applied between dense layers, followed by optional dropout regularization to reduce overfitting. The final output layer consists of two neurons for binary classification. A softmax operation converts the logits into posterior probabilities. The predicted class is the one with the higher probability score. This stage performs the final decision-making after the network has already aggregated visual, residual, and spectral cues from the previous branches.

Loss Function and Optimization

The proposed model is trained as a supervised binary classifier using cross-entropy loss. Since deepfake datasets may not always contain perfectly balanced numbers of real and fake samples, the loss function is modified using class weights. Class imbalance is handled using weighted loss. This ensures that the model does not become biased toward the majority class and that misclassification of underrepresented samples receives a sufficient penalty during optimization.

The weighted cross-entropy loss is defined as:

$$L = -\sum_{c=1}^{2} w_c y_c \log(\hat{y}_c)$$

where wc denotes the weight assigned to class c, yc is the ground-truth table, and ŷc is the predicted probability of class c.

In addition to weighted loss, label smoothing is used to prevent the model from becoming overly confident in its predictions. This improves calibration and can contribute to better generalization, especially when artifacts vary in strength across samples. The optimization objective is therefore not only to maximize training accuracy but also to encourage stable learning in the presence of noisy and heterogeneous visual evidence. This is important because benchmark studies have shown that deepfake detectors can easily overfit to dataset-specific signals and fail under domain shift [7], [8].

Training Setup

The model is trained using the Adam optimizer, which is widely used in deep learning due to its adaptive learning-rate mechanism and stable convergence behavior. The initial learning rate is set to [fill value], the batch size is [fill value], and the number of training epochs is [fill value]. These values can be adjusted depending on dataset size, GPU memory, and convergence behavior.

The implementation is developed in PyTorch, with supporting libraries for image transformation, numerical processing, and evaluation. Training can be performed on either CPU or GPU, but GPU acceleration is strongly preferred because the multi-branch design involves parallel feature extraction and repeated preprocessing operations. The experiments are conducted on [fill GPU model / CPU details].

During training, the augmented training images are fed into the three-branch model, and the optimization process updates all learnable parameters jointly in an end-to-end manner. The attention fusion weights are also learned during this process, enabling the network to discover which feature domain contributes most strongly to deepfake detection. Validation monitoring may be used to select the best-performing checkpoint and reduce overfitting. Such careful training is necessary because existing studies have shown that performance on benchmark datasets does not always translate into robust performance on unseen data [1], [7], [8]

IV. EXPERIMENTAL SETUP

Dataset Description

The experiments were conducted on the FaceForensics++ dataset, which is a widely used benchmark for manipulated facial image and video detection. In this work, the detection task is formulated as a binary classification problem with two classes: original and deepfake. Instead of directly processing full videos, frames were extracted from the videos to construct an image-level dataset for model training and evaluation. According to the final experimental report, the dataset split used a 70:30 train-test ratio with 94,686 training samples and 40,590 test samples, resulting in a total of 135,276 extracted frame samples. The split was created using a fixed random seed of 42, ensuring reproducibility of the experiments.

Implementation Details

The proposed framework was implemented in PyTorch, with support from commonly used scientific and computer vision libraries such as NumPy, OpenCV, Torchvision, and Scikit-learn. The spatial branch employed EfficientNet-B0 as a pretrained backbone, while custom preprocessing modules were used for SRM-based residual extraction in the noise branch and FFT-based spectrum generation in the frequency branch. The final report confirms the dataset configuration, class distribution, confidence statistics, and evaluation outputs; however, it does not explicitly specify the hardware platform or the exact GPU/CPU model used during training and inference. Therefore, the experiment can only be described as having been executed in the configured deep learning environment used for the study.

V. EVALUATION METRICS

The performance of the proposed model was evaluated using several standard classification metrics, including Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, Balanced Accuracy, and the Confusion Matrix. These metrics provide a comprehensive understanding of the model's effectiveness, especially in scenarios involving class imbalance.

Accuracy measures the overall proportion of correctly classified samples among all test samples and is defined as:

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

Precision quantifies the proportion of samples predicted as deepfake that are actually deepfake, indicating the reliability of positive predictions:

$$Precision = \frac{TP}{TP - FP}$$

Recall, also known as Sensitivity, measures the proportion of actual deepfake samples that are correctly identified by the model:

$$Recall(Sensitivity) = \frac{TP}{TP + FN}$$

Specificity evaluates the model's ability to correctly identify original (non-deepfake) samples:

$$Specificity = \frac{TP}{TN + FP}$$

F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics:

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall}$$

Balanced Accuracy is particularly useful when dealing with imbalanced datasets, as it considers both sensitivity and specificity equally:

$$Balanced\ Accuracy = \frac{sensitivity + specificity}{2}$$

In addition to these scalar metrics, a Confusion Matrix was used to summarize the number of true positives, true negatives, false positives, and false negatives. This gives a clearer class-wise view of the model's strengths and weaknesses. The final report also includes macro and weighted variants of precision, recall, and F1-score, which are useful for understanding performance under class imbalance.

VI. RESULTS

Quantitative Results

Based on the final report, the proposed multi-branch framework achieved an overall test accuracy of 63.75%. Although this is lower than the initially expected performance, the model still shows meaningful detection capability, particularly in identifying deepfake samples. The reported balanced accuracy of 70.95% indicates that the model performs better than the raw accuracy alone suggests, especially because the test set is imbalanced. For the deepfake class, the model achieved 95.17% precision, 60.95% recall, and an F1-score of 74.31%. The model also achieved 80.95% specificity for the original class. These results suggest that the model is highly confident when predicting deepfakes, but it misses a considerable number of fake samples and also shows uneven behavior across the two classes.

Table 1. Quantitative performance

Metric	Value
Accuracy	63.75%
Precision	95.17%
Recall / Sensitivity	60.95%
Specificity	80.95%
F1-score	74.31%
Balanced Accuracy	70.95%

Note: The above precision, recall, and F1-score correspond to the deepfake class as the positive class, consistent with the final report.

Confusion Matrix

The confusion matrix shows how the model classified the two classes in the test set. Since the report defines deepfake as the positive class and original as the negative class, the confusion matrix can be interpreted as follows: true positives correspond to correctly detected deepfakes, false negatives correspond to deepfakes misclassified as original, false positives correspond to original samples

misclassified as deepfake, and true negatives correspond to correctly identified original samples.

Table 2. Confusion matrix

Actual / Predicted	Real	Fake
Real	4590	1080
Fake	13635	21285

Class-wise Performance

The class-wise results reveal a strong imbalance in model behavior. For the deepfake class, the model achieved a precision of 95.17%, recall of 60.95%, and F1-score of 74.31% over 34,920 samples. This means that when the model predicts a sample as deepfake, it is usually correct, but it fails to detect a substantial number of actual deepfakes. For the original class, the model achieved a precision of 25.19%, recall of 80.95%, and F1-score of 38.42% over 5,670 samples. This indicates that the model is relatively good at recovering original samples when evaluated from the class-recall perspective, but predictions involving the original class are much less precise. Overall, the results suggest that the detector is more reliable for confirming deepfakes than for maintaining balanced performance across both classes.

VII. DISCUSSION

Observations

The results show that the proposed multi-branch architecture is capable of learning meaningful forensic cues from spatial, noise, and frequency-domain inputs. In particular, the very high deepfake precision of 95.17% indicates that the fusion strategy helps the model identify fake samples with strong confidence. However, the overall accuracy of 63.75% and recall of 60.95% show that the detector still misses a large portion of manipulated samples. The balanced accuracy of 70.95% is a more informative indicator here, as it reflects the unequal class distribution and shows that the model retains moderate discrimination ability despite the imbalance.

Limitations

The final report highlights several practical limitations through its evaluation statistics. First, the model does not yet achieve balanced class performance, as shown by the large gap between deepfake-class precision and original-class precision. Second, the test set itself is imbalanced, with 34,920 deepfake samples and 5,670 original samples, which can bias the learning process and distort raw accuracy interpretation. Third, the confusion matrix shows that 13,635 deepfake samples were incorrectly classified as original, indicating that the detector still struggles with a substantial subset of fake inputs. In a broader research context, deepfake detectors are also known to suffer from domain shift and reduced performance on unseen datasets or manipulation methods, especially when training data lacks sufficient diversity.

Reason Analysis

Several factors may explain the reported performance pattern. The first is class imbalance, which is evident from the support values in the final report and may influence the model to learn stronger evidence for the majority class distribution. The second is the likely presence of intra-dataset bias, where the model captures manipulation patterns specific to the training samples but fails to fully generalize to more subtle or difficult cases in the test set. The third factor is domain complexity: although spatial, noise, and frequency features are complementary, their fusion does not automatically guarantee uniformly strong performance unless the model is sufficiently trained and calibrated. More generally, prior work has shown that deepfake detectors often degrade under domain shift, especially when the training data lacks diversity in compression level, synthesis quality, facial pose, and video characteristics.

VIII. CONCLUSION

This study presented a multi-branch deepfake detection framework that combines spatial, noise-based, and frequency-domain representations through an attention-based fusion mechanism. The spatial branch captures visible facial inconsistencies,

the noise branch extracts residual forensic traces using SRM filtering, and the frequency branch identifies spectral irregularities through FFT-based analysis. By integrating these complementary cues, the framework aims to improve the robustness of deepfake classification beyond conventional single-branch detectors.

Based on the final experimental report, the proposed approach achieved an overall accuracy of 63.75% and a balanced accuracy of 70.95% on the test set. The model showed particularly strong deepfake precision of 95.17%, indicating that its positive predictions are highly reliable. However, the lower recall and the confusion matrix analysis reveal that many fake samples are still missed, which limits the overall effectiveness of the system in its current form. In future work, the framework can be improved through better class balancing, cross-dataset evaluation, more diverse training data, stronger calibration of the fusion module, and optimization for real-time deployment. These directions may help improve both generalization and class-wise consistency in practical deepfake detection scenarios.

REFERENCES

1. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019. Available: https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf
2. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018. Available: https://hal.science/hal-01867298/file/afchar_WIFS_2018.pdf
3. Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues," in Proc. European Conf. Computer Vision (ECCV), 2020. Available: https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123570086.pdf
4. I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch Recurrent Network for Isolating Deepfakes in Videos," in Proc. European Conf. Computer Vision (ECCV), 2020. Available: https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123520647.pdf
5. A. Agarwal, A. Agarwal, S. Sinha, M. Vatsa, and R. Singh, "MD-CSDNetwork: Multi-Domain Cross Stitched Network for Deepfake Detection," arXiv preprint arXiv:2109.07311, 2021. Available: https://arxiv.org/pdf/2109.07311
6. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. Machine Learning (ICML), 2019. Available: https://arxiv.org/pdf/1905.11946
7. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf
8. A. V. Nadimpalli and A. Rattani, "On Improving Cross-dataset Generalization of Deepfake Detectors," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2022. Available: https://openaccess.thecvf.com/content/CVPR2022W/WMF/papers/Nadimpalli_On_Improving_Cross-Dataset_Generalization_of_Deepfake_Detectors_CVPRW_2022_paper.pdf
9. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning Rich Features for Image Manipulation Detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018. Available:
10. B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv preprint arXiv:2006.07397, 2020. Available: