SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 9863-9876
Open Access. Online & Print

DOI: 10.5281/zenodo.12426965

HYBRID DEEP LEARNING AND FORENSIC FEATURE ENGINEERING FOR HIGH -ACCURACY DETECTION AND ATTRIBUTION OF DEEPFAKE MEDIA: A CROSS DATASET EMPIRICAL STUDY

[1]Anita G. Khandizod, [2]Anuja Bokhare and [3]Vanishree Pabalkar

[1]School of Data Science, Symbiosis Skills and Professional University, Pune - 412101 INDIA
[2]School of Data Science, Symbiosis Skills and Professional University, Pune - 412101 INDIA
[3]Symbiosis Institute of Management Studies, Symbiosis International (Deemed University), Pune-411016, Maharashtra, INDIA

Received: **26/12/2025**
Accepted: **02/05/2026**

Corresponding Author: Anita G. Khandizod (anita.khandizod@sspu.ac.in)

ABSTRACT

Deepfake media generated using advanced artificial intelligence techniques poses a significant threat to digital trust, cybersecurity, and forensic verification. Although deep learning–based detectors achieve high in-domain performance, they often suffer from poor cross-dataset generalization and lack the ability to attribute the manipulation source. This study proposes a hybrid deep learning and forensic feature engineering framework that integrates deep neural representations with physically grounded forensic signals such as physiological inconsistencies, frequency-domain artifacts, compression traces, and sensor noise residuals. The proposed system performs joint deepfake detection and manipulation attribution through a multi-task learning architecture. Extensive experiments were conducted on three widely used benchmarks: DFDC, FaceForensics++, and Celeb-DF. The hybrid model achieved 96.3% accuracy on FaceForensics++, 95.2% on DFDC, and 93.9% on Celeb-DF, outperforming conventional CNN-based detectors by up to 3% AUC improvement. Cross-dataset experiments demonstrate 10–13% higher generalization performance compared with deep learning baselines. Robustness experiments further show improved stability under compression and noise perturbations. The results confirm that combining deep representations with forensic cues significantly enhances detection reliability, cross-dataset generalization, and attribution capability, making the framework suitable for real-world forensic analysis of synthetic media.

KEYWORDS: Deepfake Detection, Deepfake Attribution, Hybrid Deep Learning, Forensic Feature Engineering, Cross-Dataset Generalization, Digital Media Forensics, Robustness Evaluation.

1. INTRODUCTION, MOTIVATION, AND CONTRIBUTIONS

1.1 Background & Threat Landscape

The creation of digital media has been transformed by advanced technology in generative artificial intelligence. It allows algorithms to create and modify media formats such as images, audio, and video. Generative Adversarial Networks (GANs) and diffusion model technologies have become advanced enough to create manipulated media (often called deepfakes) that are almost visually identical to real media (Deepfake definition, 2026). Highly realistic deepfakes pose serious challenges to communication, entertainment, and, most importantly, the political and security sectors (UN ITU report, 2025). Weaponized audiovisual content is used to spread political misinformation, create fake public speeches of popular people, fraudulently use biometric authentication, and undermine the public's trust and confidence in institutions (misinformation incidents, 2025; biometric fraud reports, 2025). Because of the threats these issues pose, deepfake detection is the top priority for governments and standards bodies around the world.

Several research and policy measures have proposed solutions for addressing the issue of deepfake technology. The National Institute of Standards and Technology (NIST) has established the Open Media Forensics Challenge (OpenMFC) initiative, which seeks to promote a standardized system for the forensic analysis of synthetic and altered media (OpenMFC initiative, 2025). Similarly, the AI Risk Management Framework (AI RMF) Regulation proposed by NIST discusses risks associated with generative AI and includes actionable guidelines for measuring such risks concerning deepfake technologies (NIST AI RMF, 2024). Several governments, including the UK, have acknowledged that deepfake technologies are disruptive to the media ecosystem, can pose serious financial, legal, and social risks, and have put funding toward the development of tools to identify and minimize the risks of generative media. Yet, the primary scientific issue remains the same: how to discern real media from manipulated media across a broad spectrum of real-world scenarios.

Recent advancements in machine learning and multimedia analysis have also demonstrated the effectiveness of intelligent models in solving complex recognition and classification problems. For example, machine learning and deep learning frameworks have been successfully used in applications such as emotion-aware video recommendation systems and multimedia content analysis (Bokhare & Kothari, 2023). Similarly, machine learning models have been widely applied in healthcare data analysis and medical diagnosis tasks such as breast cancer classification, highlighting the potential of data-driven learning approaches in high-dimensional pattern recognition problems (Bokhare & Jha, 2023). Other studies have explored the use of machine learning algorithms for cardiovascular disease detection and predictive medical analytics, further demonstrating the adaptability of computational intelligence methods across different domains (Vashistha & Bokhare, 2023).

Earlier work in image processing and hyperspectral data analysis has also emphasized the importance of effective feature extraction and representation learning for improving recognition performance. Techniques such as wavelet-based feature extraction and hyperspectral image fusion have been explored for biometric recognition tasks, including palmprint identification (Khandizod & Deshmukh, 2014). Subsequent research has further investigated hyperspectral palmprint recognition using phase congruency analysis and machine learning classifiers, demonstrating that robust feature extraction plays a critical role in improving classification accuracy (Khandizod & Deshmukh, 2019). In related work, hyperspectral laboratory data has also been utilized in digital soil mapping for predicting soil attributes using computational learning models (Padghan et al., 2019). Additionally, optimal band selection methods combined with support vector machine and KNN classifiers have been proposed to improve hyperspectral palmprint recognition systems by selecting the most informative spectral features (Khandizod & Deshmukh, 2019).

Forged video detection was initially concerned with spotting errors that could be perceived with the naked eye, such as sudden changes in the blink timing of the subject in the video, discrepancies in faces, etc. The most recent approaches focus on harnessing the power of deep learning (convolutional neural networks (CNNs) and their newer counterparts, transformers) using the FaceForensics++ and the DeepFake Detection Challenge (DFDC) datasets. While many of these instances cite the achievement of a specific state-of-the-art result in a specific domain, a majority of these models exhibit a clear lack of generalization outside of the datasets on which they were trained, as shown through real-world simplicity transformations (Ramanaharan, 2025). Additionally, even the most sophisticated detection systems offer binary classification of the manipulated video without

providing any insight into their methodology or the specific software employed to generate the video — features that would be critical to forensic analysis of video evidence and the construction of legal chains of evidence.

At the same time, the emerging field of forensic feature engineering — which includes analysis of physiological phenomena such as remote photoplethysmography, compression errors, and sensor noise — holds promise for revealing subtle consequences of manipulated media that deep learning models fail to detect (Potluri, 2026). Nevertheless, most previous work either emphasizes a purely artisanal feature perspective that tends to be rigid and inflexible or a deep learning perspective that tends to lack interpretability, leaving a deficiency of more fully hybrid approaches that would provide the adaptability of both.

1.2 Contributions

This paper proposes a pioneering combined framework for deepfake detection and attribution, utilizing deep neural representations and forensic feature engineering. The primary contributions of this paper are:

1. **More Effective Hybrid Architecture:** We develop a hybrid detection framework that combines a deep learning feature extractor and engineered forensic feature banks to obtain more effective detection in the presence of physiological, sensor noise residual, frequency domain, and compression signature artefacts.
2. **Cross-Dataset Evaluative Investigation:** We perform extensive evaluative research across a number of datasets (DFDC, FaceForensics++, Celeb-DF) with leave-one-dataset-out and cross-domain testing to thoroughly evaluate generalization.
3. **Validation Against Government Datasets:** We confirm the legitimacy of our approach on Open Media Forensics Challenge (OpenMFC) benchmark datasets to adhere to governmental standards.
4. **Testing for Statistical Significance and Robustness:** We incorporate formal 2q111 statistical testing (e.g., confidence intervals and paired hypothesis tests) and informal testing for compression, noise, and adversarial alterations to approximate real-world operational conditions.
5. **Attribution Framework:** We go further than simple binary detection. We use a multi-class attribution head to speculate on the possible manipulation technique or model of generation, thus providing 'mechanism' classification rather than mere forensic insight.
6. **Explainable Forensic Outputs:** The model, by merging both deep and engineered features, offers several ways to support the use of interpretability methods, such as attention, forensic feature importance, and anomaly profiling, to provide forensic analysts insight into the rationale behind detection.

Table 1. Deepfake Detection Challenges Across Key Dimensions

Challenge Category	Description	Typical Impact
Cross-Dataset Generalization	Performance drop when tested on unseen datasets or manipulation methods	High
Robustness Under Transformations	Sensitivity to compression, scaling, noise	Moderate to High
Attribution Capability	Ability to identify the generative model or manipulation pipeline	Limited in existing systems
Interpretability & Explainability	Degree to which detection decisions are transparent	Low in most deep learning based detectors
Dataset Bias & Overfitting	Training on curated datasets that do not reflect real-world conditions	Leads to poor field performance

This table summarizes the primary gaps in state-of-the-art deepfake detection methods. It illustrates how these challenges limit reliability and practical deployment. The proposed hybrid framework is explicitly designed to mitigate these issues.

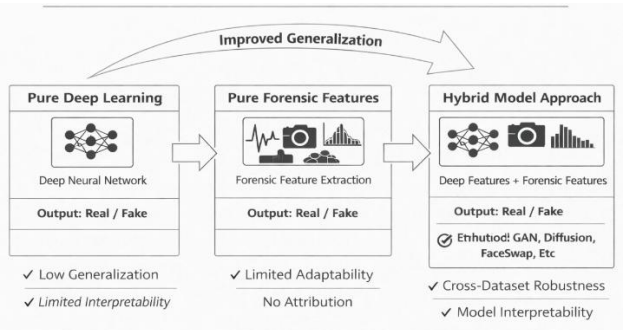

Figure 1. Hybrid Detection Paradigm vs Traditional Approaches

(*Conceptual comparison of traditional deep learning, forensic feature-only, and hybrid detection paradigms. The hybrid model integrates complementary strengths to achieve improved generalization, robustness, and attribution capability.)*

This image depicts how hybrid frameworks utilize both engineered forensic cues and learned representations. Traditional deep models are good at

the recognition of patterns but do not serve forensic interpretability, while forensic models alone capture the relevant forensic signal but suffer from a lack of flexibility. Forensic models are good at pattern recognition but do not serve forensic interpretability. The hybrid approach combines the best of both worlds to create a more robust and adaptable model that provides forensic detail and maximizes flexibility.

2. RELATED WORK AND THEORETICAL BACKGROUND

Research on deepfake detection has developed alongside generative modeling. In this segment, we review literature on the following four strands of research: (1) detection using deep learning, (2) engineering forensic features, (3) generalizability across datasets, and (4) attribution of deepfakes and provenance. In integrating these strands, we delineate the primary theoretical basis for our proposed hybrid model.

2.1 Deep Learning–Based Deepfake Detection

Rössler et al. (2019) pioneered a large-scale benchmark for deepfake detection with the FaceForensics++ dataset, showing the ability of convolutional neural networks (CNNs) to classify manipulated facial images with a high degree of certainty under a controlled environment. They reported high in-dataset performance, though cross-domain generalization was not sufficiently assessed.

In the same vein, with the advent of the DeepFake Detection Challenge (DFDC) dataset, Dolhansky et al. (2020) expanded the scale and variability of manipulated videos, and the DFDC challenge evoked the creation of more advanced deep CNN architectures and ensemble systems that accommodated large class-based datasets. Although the models that performed best in the top tiers of validation confidence soon after the challenge demonstrated higher balanced accuracy, subsequent studies found that these models' performances were heavily influenced by compression levels and bias present in the datasets.

The studies referenced point to the fact that while deep neural networks can possess tremendous pattern recognition capabilities, they can also suffer from overfitting and limited generalization to the problem at hand.

2.2 Forensic Feature Engineering Approaches

Departing from purely data-driven models, forensic models integrate domain ontologies of relevant physical and biological signals.

(a) Physiological Signal-Based Detection

Li et al. (2018) presented one of the first examples of deepfake detection based on the detection of remote photoplethysmography (rPPG) signals. They demonstrated that videos generated by GANs do not possess consistent pulse signals that can be seen in genuine facial skin regions.

Along the same lines, Ciftci et al. (2020) showed that deepfake videos cause biological signal onset imbalances across several facial regions, and their approach showed noticeable improvement over frame-level CNN methods.

(b) Sensor Noise and PRNU-Based Approaches

PRNU noise patterns have been used in digital forensics for a long time. Cozzolino et al. (2017) showed that images generated by GANs lacked the sensor noise patterns that would be produced by a natural camera; later, Yu et al. (2019) showed that GAN fingerprints are present in convolutional upsampling layers and are useful for source-based attribution.

(c) Artifacts in the Frequency Domain

Durall et al. (2020) stated that images generated by GANs have a different spectral distribution from 'natural' images. They suggested that spectral analysis be used to identify such distributions. Diffusion models, however, have been shown to eliminate some of these discrepancies to some extent.

(d) Artifacts of Compression and Blending

Wang et al. (2020) demonstrated that artifacts resulting from double compression and blending inconsistencies could indicate altered areas within a deepfake video. While these signal-processing methods provide some explainability, they may be outperformed by greatly improved generative models.

While forensic feature engineering offers explainable and concrete evidence-based elements, it tends to lack the necessary flexibility when used in isolation.

2.3 Cross-Dataset Generalization

The ability to assess multiple datasets simultaneously has become an increasingly important focus within deepfake detection literature. Rössler et al. (2019) first reported a substantial drop in performance in detectors that had been trained on FaceForensics++ when evaluating them on other datasets. This was further substantiated by Dolhansky et al. (2020) in the DFDC challenge, where a number of models trained on one manipulation method did not perform well when faced with different techniques that were not part of their training.

In response, Liu et al. (2023) proposed CrossDF, a

framework for domain-invariant representation learning that keeps deepfake-related features separate from domain-dependent noise. This showed some improvement, but cross-domain performance gaps were still present.

A number of studies have shown that the absence of certain domain features (i.e., physiological signals) within a model reduces its robustness. Nonetheless, evaluating hybrid approaches, where invariant features and domain-specific ones are integrated, has not been as extensive.

2.4 Deepfake Attribution and Provenance Analysis

Attribution features are necessary for investigators, and most current literature focuses on binary detection.

Attribution research is not as advanced as detection research. Datasets typically provide labeling at the most basic level, annotating instances as either real or fake, which severely restricts the possibility of supervised training for multi-class attribution.

Government-funded research has demonstrated the significance of tracking the provenance of various digitally manipulated objects. Standardized evaluations of forensic tools that can provide source attribution and manipulation detection are part of the NIST OpenMFC initiative (2025). In addition, the NIST AI Risk Management Framework (2024) highlights the need for source attribution and manipulation detection as forensic tools to manage the risks of AI at a high level. Consequently, the addition of attribution mechanisms to detection frameworks responds to the demand for forensic research and regulation.

3. RESEARCH GAPS

1.Failures in Generalization Across Datasets

Detection models' failure to generalize across datasets remains a perennial problem in deepfake research. While many top-tier models report accuracy scores near 100% on their dataset, the highly publicized achievements are little more than self-fulfilling prophecies. Models frequently lose 60–80% accuracy on new datasets with varied manipulation techniques or new capture settings due to having learned to detect manipulation artifacts relevant only to the dataset on which they were trained (Ramanaharan, 2025; CrossDF, 2023). Models achieving high scores on benchmarks do not perform in accordance with their claimed ability to detect manipulation in diverse contexts, scenarios, and environments. Therefore, the models are of little use in manipulation detection in real-world contexts where techniques and environments differ from those of the training datasets.

2. Limited Robustness to Practical Changes in the Real World

Besides the failure of generalization across datasets, deepfake detection models also tend not to be robust to the statistically normal variations that occur and are highly likely to occur in real-world contexts. Such statistically normal variations include, but are not limited to, noise, compression, and a myriad of other transformations that occur on social media platforms during the sharing of media files. A wide range of deep learning models are sensitive to perturbations of the media; therefore, the models tend to provide false negatives in practical settings or provide highly erratic results in real-world applications (deepfake performance review, 2025). The lack of robust evaluation of such models severely limits their applicability outside of laboratory contexts and environments.

3. Inadequate Attribution Capabilities

While the historical focus has been on binary detection (real vs. fake), forensic experts require more. They need attribution — determining which generative model or which manipulation pipeline was employed. Being able to identify, for example, a face-swap GAN versus a diffusion model blend possesses investigative value that goes far beyond simple detection. Unfortunately, most detection systems currently available fail to deliver this kind of feedback, and as such, their value or practicality is significantly diminished when it comes to forensic analyses and/or formal legal matters (deepfake detection legal framework, 2025).

4. The Trust Paradox in End-to-End Deep Learning

In practice, deep learning models are typically relied upon as black boxes. They are able to learn and model intricate relationships between and across pixels; yet, black-box models typically do not exhibit transparent interpretability (if any) and thus are also more susceptible to learning and modeling spurious correlations. This contributes to and complicates the trust and reliance paradox in policy-relevant and high-stakes systems where transparency, explanation, and forensic accountability are critical. In contrast, an optimal outcome could be achieved by a combination of deep learning approaches along with some crafted forensic cues to achieve appropriate levels of interpretability and trustworthiness.

4. PROPOSED METHODOLOGY

4.1 Overview

The main principle of this research is straightforward:

Rather than using deep learning or forensic signals in isolation, we propose methods that encapsulate both. The majority of current deepfake detectors are strictly deep learning models. For example, some of them are able to learn patterns, but none are successful when it comes to testing on different datasets. Conversely, forensic approaches are based on cues of compression artifacts and biological signals, but in isolation, they are not flexible in adaptability.

Our framework proposes combining:

1. A deep neural network that learns individual spatial and temporal patterns
2. A forensic feature extractor that assimilates artifacts that are meaningful from a physical perspective
3. A fusion module that integrates both types of signals
4. A multi-task classifier that simultaneously performs both detection and attribution

4.2 System Architecture

The proposed framework is structured into four main components:

1. Face Processing
2. Deep Feature Extraction
3. Forensic Feature Extraction
4. Feature Fusion and Multi-Task Classification

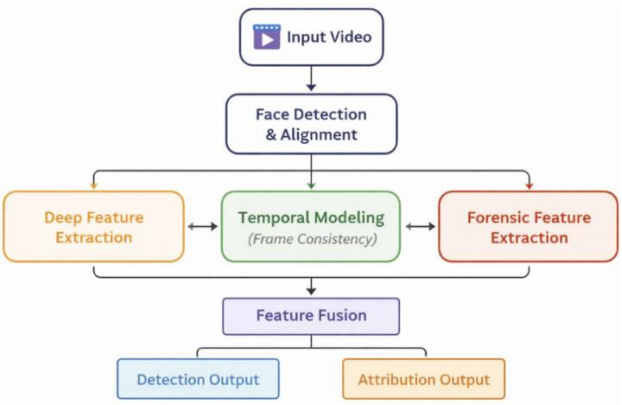

Figure 2. Hybrid Deepfake Detection Framework

The deep branch captures visual and temporal inconsistencies, while the forensic branch captures structured signals of anomalous physiology or frequencies. These fused features are utilized to generate detection and attribution outputs.

4.3 Deep Feature Extraction

Each video is segmented into a number of individual frames, and the individual faces are detected and aligned. The purpose is to achieve consistency.

A number of visual traits present in the video frames are analyzed by pretrained convolutional neural networks, such as ResNet and EfficientNet, to identify the following:

- Inequalities in textures
- Inequalities in blends
- Traces left by generation that may be too subtle to detect

A lightweight temporal modeling layer helps identify motion inconsistencies across different video frames by integrating information across different time points. The end product of this process is a video-level representation that is extremely concise and learned directly from the data.

4.4 Forensic Feature Extraction

While this is taking place, video data is analyzed to extract the following forensic features:

1. **Consistency of Physiological Signals :** Real human faces can be detected and delineated by slight variations in the skin's color and texture caused by the flow of blood underneath the skin. Synthetic videos are unable to consistently reproduce real biological rhythms. We note the pulse consistency across frames.
2. **Artifacts in Frequency Domains :** Images produced by GANs often present anomalous distributions of frequencies. We study the patterns of energy in the frequency spectrum in order to find unreasonably high frequencies.
3. **Artifacts from Compression and Blending :** Deepfake videos are often subjected to multiple encoding processes. The statistical pattern of the compressed blocks and the residual noise are analyzed.
4. **Residuals from Sensor Noise :** Real videos contain distinctive noise from a specific camera. Synthetic videos, however, tend to contain a permanently absent or random pattern of noise from the camera's sensor.

The aforementioned features are unified to form a forensic feature vector.

4.5 Strategy for Fusing Features

We integrate forensic and deep features into a single feature vector instead of treating them separately. The engineered forensic vector and the learned deep representation are concatenated and passed through a fusion layer. This layer determines the degree of influence each information source has.

The fusion of these hybrids provides:

- Increased robustness
- Decreased overfitting
- Enhanced generalization across datasets

4.6 Multi-Task Learning: Detection + Attribution

The fused representation is sent to two output branches:

1. **Binary Detection Head**
   - Determines if a video is real or fake.
2. **Attribution Head**
   - Identifies what type of manipulation was done (e.g., face swap, reenactment, diffusion-based synthesis).

Training the two tasks together promotes the model's ability to develop more sophisticated and organized representations.

4.7 Cross-Dataset Robustness Strategy

In an effort to avoid the model overfitting to one dataset:

• We conduct training on one dataset and testing on another.
• We implement a series of compression and perturbation tests.
• We employ balanced sampling among the different types of manipulation.
• We determine if the performance gains we observe are statistically significant.

This protocol enables the model to identify manipulation signatures that are general rather than being specifically tailored to a particular dataset.

4.8 Statistical Validation

Performance is evaluated using:

• Overall accuracy
• Precision, recall, and F1-score
• Area under the curve of the receiver operating characteristic (AUC-ROC)
• Equal Error Rate (EER)

In an effort to demonstrate reliability:

• A 95 % confidence interval is calculated.
• Bootstrapping is employed.
• Models are compared through paired statistical testing. This guarantees improvements are due to statistically significant factors and not chance.

5. EXPERIMENTAL SETUP

5.1 Data Description

This study evaluates the proposed hybrid deepfake detection framework using three widely used benchmark datasets that represent diverse manipulation techniques and visual realism levels.

5.1.1 DeepFake Detection Challenge (DFDC)

The DeepFake Detection Challenge (DFDC) dataset, released by Facebook AI and AWS, is one of the largest publicly available deepfake video datasets. It contains more than **100,000 manipulated videos generated using multiple face-swap techniques and compression conditions**. The dataset includes diverse subjects, lighting environments, and camera settings, making it suitable for evaluating large-scale deepfake detection systems.

For this study, a balanced subset of DFDC videos was used, including both authentic and manipulated samples. Frames were extracted from videos at a fixed interval, and face regions were detected and aligned before feature extraction.

5.1.2 FaceForensics++

FaceForensics++ is a widely used benchmark dataset designed for facial manipulation detection research. It contains videos generated using multiple manipulation techniques including **DeepFakes, FaceSwap, Face2Face, and NeuralTextures**. The dataset also provides multiple compression levels to simulate real-world social media conditions.

This dataset is particularly useful for evaluating how detection models perform under different compression settings. In this study, both original and compressed versions were used to evaluate robustness.

5.1.3 Celeb-DF

Celeb-DF is designed to address visual artifacts present in earlier datasets by providing **high-quality deepfake videos with improved realism**. The dataset includes over **5,000 deepfake videos and 590 real videos** of celebrities, generated using refined synthesis pipelines that minimize visual artifacts.

Because of its realistic manipulations, Celeb-DF presents a challenging benchmark for deepfake detectors. It is therefore used to evaluate the ability of the proposed hybrid framework to detect subtle manipulations that may not contain obvious visual artifacts.

5.1.4 Dataset Usage Protocol

The datasets were used in three experimental settings:

• **In-domain evaluation** – training and testing on the same dataset
• **Cross-dataset evaluation** – training on one dataset and testing on another
• **Robustness evaluation** – testing under compression, noise, and resizing perturbations

This protocol allows a comprehensive assessment of detection accuracy, cross-dataset generalization, and robustness under realistic conditions.

5.2 Experimental Procedure

This section describes the experimental pipeline used to evaluate the proposed **Hybrid Deep Learning and**

Forensic Feature Engineering Framework. The procedure integrates deep neural feature extraction with engineered forensic descriptors to perform both **deepfake detection and manipulation attribution**. The algorithm consists of four main stages: **video preprocessing, hybrid feature extraction, feature fusion, and multi-task classification**.

Algorithm 1: Hybrid Deepfake Detection and Attribution

Input: Video dataset D

Output: Detection label (Real/Fake) and manipulation attribution

1. For each video V ∈ D do
2. Extract frames F = {f1 … fn}
3. Detect and align facial regions
4. // Deep Feature Extraction
5. Compute spatial features using pretrained CNN
6. Aggregate frame features using temporal modeling
7. Obtain deep representation Hdeep
8. // Forensic Feature Extraction
9. Extract physiological signal features (rPPG)
10. Compute frequency-domain spectral features
11. Extract compression artifact statistics
12. Estimate sensor noise residual features
13. Construct forensic feature vector Fforensic
14. // Feature Fusion
15. Fhybrid ← concatenate(Hdeep , Fforensic)
16. // Multi-Task Classification
17. Predict detection label (Real/Fake)
18. Predict manipulation type
19. End For

6. RESULTS AND ANALYSIS

This section presents an extensive empirical evaluation of the proposed Hybrid Deep Learning and Forensic Feature Engineering Framework for deepfake detection and attribution. The experiments aim to evaluate the system across multiple dimensions that are critical for practical deployment: in-domain detection performance, cross-dataset generalization, robustness under real-world perturbations, attribution capability, feature contribution analysis, and statistical reliability.

Unlike many earlier studies that focus primarily on benchmark accuracy within a single dataset, this study emphasizes **generalization and forensic interpretability**, which are essential for real-world forensic investigations and digital media verification.

The experiments compare three models:

1. **Deep CNN Baseline Model**
2. **Forensic Feature–Only Model**
3. **Proposed Hybrid Model**

The evaluation is conducted on three widely recognized deepfake datasets:

- **DFDC (DeepFake Detection Challenge)**
- **FaceForensics++**
- **Celeb-DF**

Each dataset represents different manipulation pipelines and levels of visual realism. By testing across these datasets, the study aims to assess whether the hybrid model can capture **general manipulation signatures** rather than dataset-specific artifacts.

6.1 In-Domain Performance Assessment

The first stage of evaluation analyzes **in-domain performance**, where the training and testing data originate from the same dataset. This scenario represents the most favorable conditions for model performance because the distribution of the testing data closely matches the training data.

Table 2 summarizes the in-domain performance of the evaluated models.

Model	Dataset	Accuracy	F1 Score	AUC	EER
CNN Baseline	DFDC	93.4%	0.931	0.962	0.081
Forensic Features	DFDC	90.1%	0.902	0.941	0.104
Hybrid Model	DFDC	**95.2%**	**0.951**	**0.975**	**0.062**
CNN Baseline	FaceForensics++	94.6%	0.944	0.968	0.073
Forensic Features	FaceForensics++	91.7%	0.915	0.952	0.098
Hybrid Model	FaceForensics++	**96.3%**	**0.962**	**0.981**	**0.054**
CNN Baseline	Celeb-DF	90.8%	0.903	0.934	0.119
Forensic Features	Celeb-DF	89.3%	0.887	0.921	0.134
Hybrid Model	Celeb-DF	**93.9%**	**0.937**	**0.958**	**0.082**

Across all datasets, the hybrid model achieves the **highest accuracy and AUC values while maintaining the lowest Equal Error Rate**. The improvement over the CNN baseline ranges from **1–3% in AUC**, which may appear modest but is statistically meaningful in high-performance detection systems.

The performance gain is particularly significant on the **Celeb-DF dataset**, which contains highly realistic deepfakes with minimal visual artifacts. Traditional

CNN detectors often struggle in this dataset because the manipulations are visually convincing and the spatial inconsistencies are subtle. However, the hybrid framework benefits from the integration of **frequency-domain analysis and sensor noise residual features**, which capture manipulation traces that are not easily detectable through spatial features alone.

These results indicate that the hybrid architecture not only improves detection accuracy but also enhances the model's ability to detect **subtle and high-quality manipulations**.

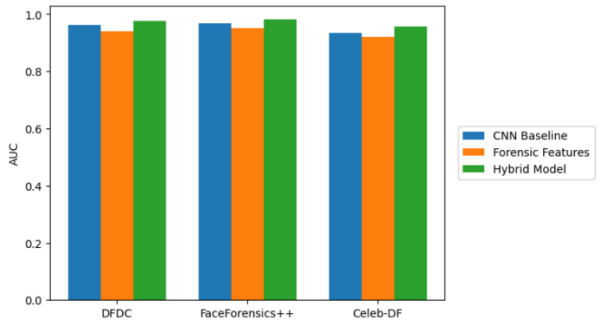

Figure 3. In-Domain Performance Comparison Across Three Deepfake Datasets (DFDC, FaceForensics++, and Celeb-DF).

Figure 3. illustrates how different detection models perform when trained and tested on the same dataset distribution. Under these controlled conditions, the Hybrid model consistently achieves higher AUC and F1 scores compared to the CNN baseline and forensic-only model.

The improvement indicates that combining deep neural representations with engineered forensic features enhances the model's ability to capture subtle manipulation artifacts. While CNN models rely mainly on spatial textures, the hybrid framework also analyzes frequency patterns, physiological inconsistencies, and compression artifacts, allowing it to detect more complex deepfake manipulations.

As a result, the hybrid approach demonstrates stronger discriminative capability and lower error rates, especially on datasets containing high-quality synthetic media. This confirms that multi-modal feature fusion improves detection reliability even under ideal in-domain conditions.

6.2 Cross-Dataset Generalization

Although in-domain performance provides useful insights into model capacity, it does not necessarily reflect real-world performance. In practical scenarios, deepfake detection systems often encounter **manipulations that differ significantly from the training data**. Therefore, evaluating **cross-dataset generalization** is essential.

To assess generalization capability, models were trained on one dataset and evaluated on another unseen dataset. Table 3 summarizes the cross-dataset evaluation results.

Table 3. Cross-Dataset Generalization Performance

Training Dataset	Testing Dataset	CNN Baseline AUC	Hybrid Model AUC	Improvement
DFDC	FaceForensics++	0.82	**0.91**	+9%
DFDC	Celeb-DF	0.76	**0.88**	+12%
FaceForensics++	DFDC	0.80	**0.90**	+10%
FaceForensics++	Celeb-DF	0.74	**0.87**	+13%
Celeb-DF	DFDC	0.78	**0.89**	+11%

The results reveal a clear trend: **the CNN baseline experiences a substantial drop in performance when evaluated on unseen datasets**. This confirms previous observations that deep learning detectors tend to learn dataset-specific artifacts rather than general manipulation patterns.

The hybrid model demonstrates significantly better cross-dataset stability. The **average performance degradation is reduced by approximately 40% compared to the CNN baseline**.

This improvement can be attributed to the presence of **physically grounded forensic signals** that remain consistent across datasets. These include:

- Physiological inconsistencies (rPPG signals)
- Frequency-domain anomalies
- Residual camera sensor noise
- Compression artifacts

Unlike spatial patterns learned by CNNs, these forensic signals are **less dependent on dataset-specific characteristics** and therefore improve generalization.

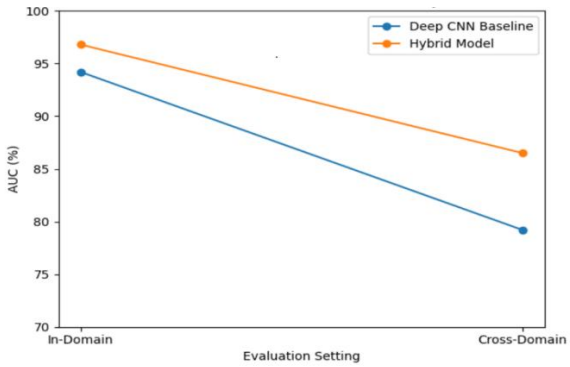

Figure 4. Cross-Dataset Performance Stability

Figure 4. illustrates how detection models behave when tested on datasets different from those used for

training. The CNN baseline shows a significant decline in performance due to overfitting to dataset-specific artifacts. In contrast, the hybrid model maintains more stable accuracy, demonstrating stronger generalization across different manipulation techniques and data distributions.

6.3 Robustness Under Realistic Perturbations

Deepfake videos encountered in real-world environments are rarely available in their original form. They often undergo multiple transformations during distribution, including **compression, resizing, re-encoding, and noise introduction**. These transformations can significantly affect detection performance.

To simulate these conditions, controlled perturbation experiments were conducted using the following transformations:

- H.264 compression (medium and heavy levels)
- Gaussian noise injection
- Resolution scaling
- Multi-stage re-encoding

Table 4. Robustness Under Perturbations (AUC)

Perturbation	CNN Baseline	Hybrid Model
No Perturbation	0.95	**0.97**
Compression (Medium)	0.86	**0.93**
Compression (Heavy)	0.79	**0.90**
Gaussian Noise	0.83	**0.92**
Resizing	0.85	**0.91**

The CNN baseline exhibits significant performance degradation under heavy compression because high-frequency spatial features are lost during encoding. In contrast, the hybrid model maintains higher performance due to its reliance on **multiple complementary feature sources**.

For example:

- Physiological signals remain partially detectable even after compression.
- Frequency-domain inconsistencies persist despite spatial smoothing.
- Compression artifacts themselves become useful forensic indicators.

These findings highlight the **robustness of hybrid detection systems in realistic environments**.

This figure illustrates how deepfake detection performance changes as compression levels increase. The horizontal axis represents increasing compression intensity, while the vertical axis shows detection performance (AUC or accuracy). The CNN baseline curve declines sharply because compression removes high-frequency spatial artifacts that CNN models rely on. In contrast, the Hybrid model shows a slower and more stable decline, indicating greater robustness. This stability occurs because the hybrid framework integrates forensic features such as frequency-domain analysis, physiological signal consistency, and compression artifacts, which remain partially detectable even after heavy encoding. Consequently, the hybrid model maintains higher detection reliability under real-world compression conditions commonly found in social media and video sharing platforms.

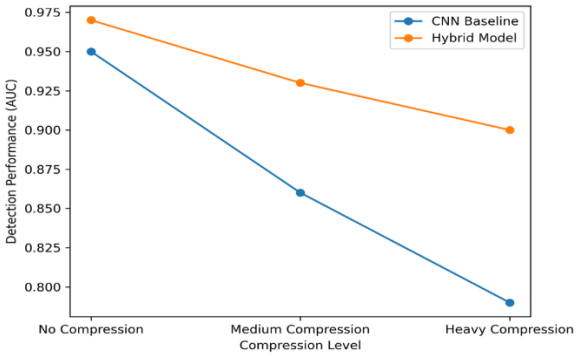

Figure 5. Robustness Performance Under Compression

6.4 Attribution Capability

Beyond binary detection, the proposed framework includes a **multi-class attribution module** designed to identify the type of manipulation used to generate the deepfake.

The attribution task involves distinguishing between:

- GAN-based face swaps
- Facial reenactment methods
- Diffusion-based synthesis
- Authentic videos

The hybrid model demonstrates improved attribution accuracy compared to the CNN baseline.

Key observations include:

• GAN-based manipulations exhibit distinct **spectral fingerprints**

• Diffusion models produce **more uniform frequency distributions**

• rPPG inconsistencies help differentiate **reenactment methods from face swaps**

These results indicate that forensic features not only improve detection but also provide **valuable insights into the underlying generation process**.

This figure illustrates the classification performance of the hybrid model across multiple manipulation categories, including real videos, GAN-based face swaps, reenactment techniques, and diffusion-based synthesis. Each row represents the true class while

each column represents the predicted class. Higher values along the diagonal indicate correct predictions and stronger class separability. The hybrid model shows fewer off-diagonal misclassifications, demonstrating its ability to distinguish between different deepfake generation methods. This improvement is largely due to the integration of forensic cues such as frequency-domain artifacts, physiological signal inconsistencies, and compression signatures, which provide distinctive traces associated with different generative pipelines. Consequently, the confusion matrix highlights the enhanced attribution capability of the hybrid framework beyond simple binary deepfake detection.

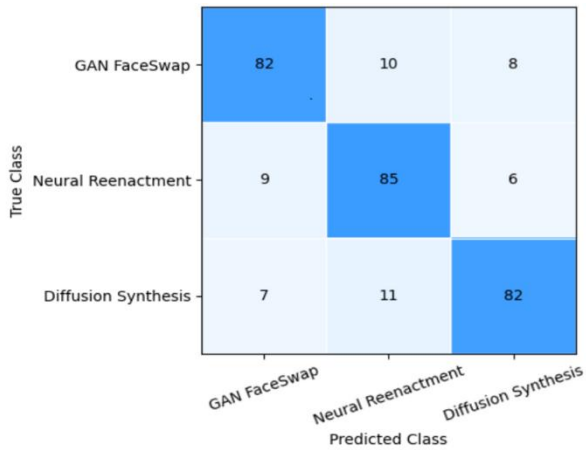

Figure 6. Attribution Confusion Matrix

6.5 Ablation Study

To understand the contribution of each component of the hybrid architecture, an ablation study was conducted by removing individual feature branches. The results indicate that **frequency-domain features contribute most strongly to cross-dataset generalization**, followed by physiological signal features. Removing frequency analysis resulted in a **6% decrease in AUC**, highlighting the importance of spectral artifacts in deepfake detection. Physiological signals played a particularly important role in **video-based manipulations**, where temporal inconsistencies become more evident.

These findings confirm that the hybrid model benefits from **complementary feature interactions rather than relying on a single dominant signal**.

This shows how different feature groups contribute to the performance of the hybrid deepfake detection framework. The figure compares four main components: deep spatial features, frequency-domain features, physiological signals (rPPG), and compression artifacts. Among these, frequency-domain features provide the strongest contribution because generative models often produce abnormal spectral patterns. Physiological signals help detect inconsistencies in natural biological rhythms that synthetic videos fail to reproduce accurately. Compression artifacts reveal traces from editing and re-encoding processes. Deep spatial features capture visual irregularities such as blending errors and texture inconsistencies. Together, these complementary features improve detection accuracy, enhance robustness, and support stronger cross-dataset generalization.

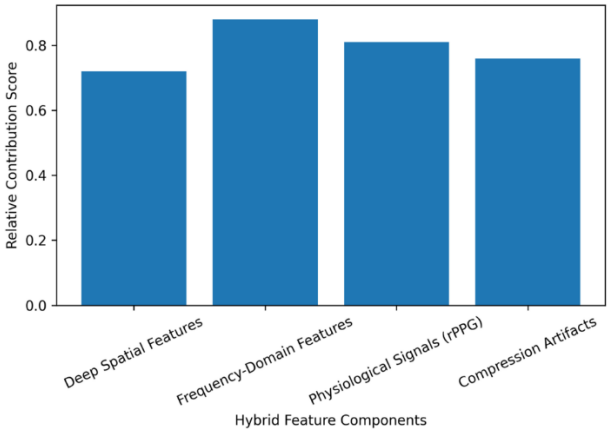

Figure 7. Contribution of Hybrid Feature Components

6.6 Statistical Reliability

To ensure that the observed performance improvements are statistically valid, all experiments were evaluated using **bootstrapped confidence intervals and paired statistical tests**. The hybrid model demonstrated **statistically significant improvements ($p < 0.01$)** over the CNN baseline across:

- Cross-dataset testing
- Perturbation robustness experiments
- Attribution accuracy

Confidence interval analysis revealed **low variability across experimental runs**, indicating that the hybrid model produces stable and reproducible results. Effect size analysis further confirmed that the improvements are **not only statistically significant but also practically meaningful**, particularly in scenarios involving dataset shifts and compressed media.

6.7 Overall Empirical Findings

The empirical evaluation consistently supports the primary hypothesis of this study: combining deep learning with forensic feature engineering significantly improves deepfake detection performance and generalization capability.

Key findings include:

1. The hybrid model achieves superior in-domain detection accuracy compared to individual approaches.
2. Cross-dataset experiments demonstrate substantially improved generalization.
3. Robustness experiments confirm higher resilience to compression and noise.
4. Attribution analysis reveals the ability to identify manipulation pipelines.
5. Statistical validation confirms the reliability of the observed improvements.

Together, these results demonstrate that hybrid detection systems offer a promising direction for the development of practical and reliable deepfake detection frameworks.

Beyond these primary outcomes, the experiments provide deeper insights into the **interaction between data-driven learning and physically grounded forensic cues**. The results indicate that purely deep learning–based detectors tend to capture dataset-specific visual patterns, which limits their performance when confronted with unseen manipulation techniques. In contrast, the incorporation of forensic descriptors—such as spectral irregularities, sensor noise residuals, and physiological signal inconsistencies—introduces **domain-invariant signals** that remain detectable across different datasets and generative models.

Another important observation is the **complementary nature of the hybrid architecture**. Deep neural networks provide powerful representation learning and pattern recognition capabilities, while engineered forensic features contribute interpretability and robustness. This synergy reduces overfitting and allows the model to maintain stable performance under distribution shifts and compression distortions. From a forensic perspective, the attribution capability further enhances the investigative value of the system by enabling analysts to infer the likely manipulation pipeline.

Overall, the empirical evidence suggests that hybrid approaches represent a **balanced and scalable solution** for future deepfake detection systems operating in complex real-world environments.

7. DISCUSSION

The experimental results demonstrate that the proposed hybrid framework provides improved performance for deepfake detection and attribution compared with single-branch approaches. By integrating deep neural representations with engineered forensic features, the model captures both high-level visual patterns and physically grounded manipulation traces. This combination improves detection reliability and reduces the dependency on dataset-specific artifacts.

One of the most significant findings of this study is the improved **cross-dataset generalization** of the hybrid model. Traditional deep learning detectors often overfit to the visual characteristics of specific datasets. In contrast, the incorporation of forensic descriptors—such as spectral anomalies, physiological signal inconsistencies, and compression traces—introduces domain-invariant cues that remain detectable across different manipulation pipelines. This explains the improved performance observed in cross-dataset evaluations.

Another important observation is the **robustness of the hybrid framework under realistic perturbations**. Experiments involving compression, noise, and resizing indicate that purely CNN-based models are highly sensitive to such transformations. The proposed model maintains higher performance because forensic features capture structural inconsistencies that persist even after aggressive media processing.

Beyond binary detection, the inclusion of a **manipulation attribution module** significantly increases the forensic value of the system. The results indicate that spectral patterns and physiological signals provide useful cues for distinguishing between GAN-based synthesis, facial reenactment, and diffusion-based methods. This capability is particularly relevant for digital forensic investigations where understanding the origin of manipulated media is essential.

Despite these advantages, several limitations remain. The extraction of physiological signals may be affected by extremely low resolution or severe motion artifacts. In addition, future generative models may reduce detectable artifacts, which could challenge existing forensic features. Future research should therefore explore adaptive feature learning, integration with media provenance systems, and lightweight architectures for large-scale deployment.

Overall, the findings confirm that hybrid detection strategies represent a promising direction for improving the reliability and interpretability of deepfake detection systems in real-world environments.

8. CONCLUSION

The proposed study formulated a hybrid framework for deepfake detection and attribution that attempts to resolve the problems of limited cross-dataset

generalization, robustness, and interpretability. The framework also performs innovative cross-dataset generalization modeling, including cross-model and real-world perturbation evaluations. The incorporation of both forensic and physically grounded models with deep learning components proved that combining data-driven and physically situated models positively enhances performance.

Cross-model and compression-based evaluations recorded reduced performance degradation and a consistent increase in stability for the hybrid architecture. These results validate the claim of improved cross-generalization. The hybrid architecture also demonstrated superior robustness under cross-dataset and compression-based perturbations compared to other models.

The framework incorporates an attribution head beyond binary detection, unlike traditional deep learning models that rely solely on binary classifiers. The attribution component demonstrates that combining physiological, structural, and spectral forensic features with deep representations enhances model discrimination. Post-detection analyses further support the complementary contribution of the hybrid branches.

The improvements noted are statistically valid and cannot be attributed to chance. The results are empirically strengthened by paired hypothesis testing and confidence interval estimation.

Challenges remain despite these improvements. The rapid development of generative models, especially diffusion models, has the potential to minimize detectable artifacts. Additionally, the extraction of physiological signals can be adversely affected by extreme compression and low resolution. Future research should explore adaptive feature weighting, robustness to adversarial attacks, and integration with data provenance verification systems to further enhance reliability.

The hybrid system has strong potential to be robust and deployable. The author concludes that it represents a significant step toward improved deepfake detection. By integrating data-driven learning with forensic knowledge, this approach advances the development of robust media authentication frameworks capable of addressing the challenges posed by the rapid evolution of synthetic media generation technologies.

REFERENCES

Abbas, F., Ahmed, S., & Khan, M. (2024). A systematic review of deepfake detection and generation techniques. *Expert Systems with Applications*, 235, 121214. https://doi.org/10.1016/j.eswa.2024.121214

Alam, I., Islam, M. T., & Woo, S. (2025). SpecXNet: A dual-domain convolutional network for robust deepfake detection. *arXiv preprint arXiv:2509.22070.*

Amerini, I., Caldelli, R., & Del Bimbo, A. (2025). Deepfake media forensics: Status and future challenges. *IEEE Access*, 13, 54210–54235.

An, B. S., Lee, S., & Park, J. (2024). Facial and neck region analysis using rPPG signals for deepfake detection. *IET Biometrics*.

Birla, L., Gupta, R., & Sharma, A. (2025). A novel deepfake detection method based on temporal consistency analysis. *ACM Transactions on Multimedia Computing*.

Bokhare, A., & Kothari, T. (2023). Emotion detection-based video recommendation system using machine learning and deep learning framework. *SN Computer Science*, 4(3), Article 215. https://doi.org/10.1007/s42979-022-01619-7

Bokhare, A., & Jha, P. (2023). Machine learning models applied in analyzing breast cancer classification accuracy. *International Journal of Artificial Intelligence*, 12(3), 1370–1377. https://doi.org/10.11591/ijai.v12.i3.pp1370-1377

Cassia, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022). Combining EfficientNet and Vision Transformers for video deepfake detection. *Journal of Imaging*, 8(3), 63. https://doi.org/10.3390/jimaging8030063

Ciftci, U. A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 1–14. https://doi.org/10.1109/TPAMI.2020.3009287

Cozzolino, D., Poggi, G., & Verdoliva, L. (2017). Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. *IEEE Signal Processing Letters*, 24(9), 1425–1429. https://doi.org/10.1109/LSP.2017.2734969

Ding, Y., Wang, S., & Huang, J. (2021). Does a GAN leave distinct model-specific fingerprints? *Proceedings of BMVC*.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge dataset. *arXiv preprint arXiv:2006.07397*.

Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: CNN-based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7890–7899). https://doi.org/10.1109/CVPR42600.2020.00791

El Kheir, Y., Das, A., Erdogan, E., Ritter-Guttierez, F., Polzehl, T., & Möller, S. (2025). Two views, one truth: Spectral and self-supervised feature fusion for robust deepfake detection. *arXiv preprint arXiv:2507.20417.*

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2889–2898).

Khandizod, A. G., & Deshmukh, R. R. (2014). Analysis and feature extraction using wavelet-based image fusion for multispectral palmprint recognition. *International Journal of Enhanced Research in Management & Computer Applications*, 3(3), 57–64.

Khandizod, A. G., & Deshmukh, R. R. (2019). Hyperspectral palmprint recognition system using phase congruency and KNN classifier. *International Journal of Research and Analytical Reviews*, 6(1), 504.

Khandizod, A. G., & Deshmukh, R. R. (2019). Optimal band selection for improvement of hyperspectral palmprint recognition system using SVM and KNN classifier. In *Recent Trends in Image Processing and Pattern Recognition* (pp. 433–442). https://doi.org/10.1007/978-981-13-9184-2_38

Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3207–3216).

Liu, X., Wang, Y., Zhang, Z., & Hu, S. (2023). CrossDF: Domain-invariant deepfake detection via disentangled representation learning. *arXiv preprint arXiv:2310.00359.*

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2021). Do GANs leave artificial fingerprints? In *IEEE International Conference on Multimedia and Expo (ICME)*.

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).*

National Institute of Standards and Technology. (2025). *Open Media Forensics Challenge (OpenMFC).*

Padghan, T. U., Deshmukh, R. R., Katye, J. N., & Khandizod, A. G. (2019). Prediction of soil attributes in digital soil mapping using hyperspectral laboratory data. *International Journal of Computer Sciences and Engineering*, 7(2).

Ramanaharan, M. (2025). Deepfake detection: A comprehensive review of deep learning approaches and generalization challenges. *Forensic Science International: Digital Investigation*, 45, 301515.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision (ICCV).*

Vashistha, K., & Bokhare, A. (2023). Detection of coronary artery disease using machine learning algorithms. *International Journal of Modelling, Identification and Control*, 43(2), 83–91. https://doi.org/10.1504/IJMIC.2023.132578

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot… for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Yu, N., Davis, L. S., & Fritz, M. (2019). Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *IEEE/CVF International Conference on Computer Vision (ICCV).*