A Hybrid GAN-CNN Pipeline for Enhanced Deepfake Detection

Tarunikka Suresh[1], Raja Muthalagu[2*]

[1,2] *Department of Computer Science and Engineering, Birla Institute of Technology and Science Pilani, Dubai Campus, Dubai, United Arab Emirates*

Abstract

Deepfake media has become a global risk due to hyper-realistic face and voice manipulation in misinformation, fraud, identity theft, and political disruption. While deep learning-based detectors are extremely accurate on curated benchmark datasets, performance collapses in real world due to cross domain shifts, compression artifacts, adversarial perturbations, and limited generalizability. The present study addresses this issue by introducing a novel hybrid deepfake detection program which combines with GAN based adversarial augmentation with XceptionNet classification to improve detection robustness. A lightweight U-Net GAN refiner produces synthetic perturbations of images with boundary inconsistencies, compression artifacts, color changes, and texture distortions, producing hard-to-identify fake images. The refined samples were added to countless real and fake images to train a robust classifier. The model produces improved performance on the Face forensics ++ dataset: 95.1% accuracy, 0.971 precision, 93.3% recall, 95.2% F1 score, and 0.984 AUC. Overall, this study evidenced that GAN based adversarial augmentation improves the decision boundaries at a CNN-based detector while compensating for overall dataset limitations. This study emphasizes that adversarial synthesis can impact modern deepfake forensics and build on future multimodal and context-aware detection systems.

***Keywords*: Deepfake Detection, GAN, XceptionNet, Adversarial Learning, U-Net, Multimedia Forensics, Representation Learning, FF+ dataset, CNN- based detector**

1. Introduction

Deepfake technology has rapidly advanced due to innovations in deep generative models, especially Generative Adversarial Networks (GANs), autoencoders, and diffusion-based models. These models can create lifelike manipulations and deeper fakes of faces, emotions, and identity markers, allowing created images and videos to be nearly indistinguishable from real images, or at least from some point of view. While deepfakes were once seen as merely a creative pastime, they became a significant, pervasive global threat that can support misinformation, undermine democracy, create opportunities for financial crimes, facilitate cybercrime, and violate individual identities. As manipulated media propagates across social media platforms, like YouTube, TikTok, and Instagram, people's means of authenticating visual media through digital communication becomes critical to prevent undermining public trust. This is particularly troubling in critical areas of journalism, legal evidence, and digital forensics that have accepted visual evidence as reliable. This reliability problem is exacerbated by true distortions: once videos are uploaded to an online platform, the videos will be resized, re-encoded, compressed, and filtered by algorithms, which will erode many of the artifact-based cues on which CNN-based detectors rely.

The importance of creating reliable deepfake detection tools emerges from continuous advances in generative models. With ongoing enhancements in architectures like Style GAN, DeepFaceLab, and diffusion transformers, as layers promise to replace many of the classical artifacts (e.g., blending inconsistencies, facial boundary mismatches, texture abnormalities) that the first detection models relied upon, we also uncover the limitations of standard CNN-based detectors, while these models may demonstrate consistent performance on publicly available

academic benchmarks, they often unravel abruptly when they explore promotional manipulations or more natural videos, both of which may be unseen. As an example, shown in Fig 1. FaceForensics++ a widely used academic benchmarking dataset of manipulated and original videos, leads to robust performer models in the benchmark based on FF++. And yet, models trained exclusively on FF++ struggle to generalize to a deepfake video, especially due to changes in lighting conditions, camera quality, facial features, manipulation pipeline, etc. The gap in domains creates fragile detectors that appear responsive in lab settings yet completely fail in applied situations. Therefore, reinforcing generalization is among the most challenging aspects of deepfake forensics.

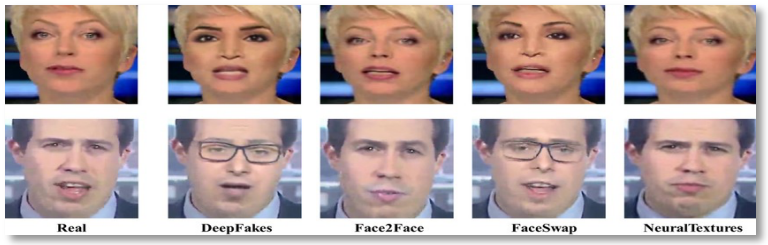

Figure 1: Examples from the FF++ dataset showing real and manipulated frames generated through Deepfakes, Face2Face, Face Swap, and Neural Textures.

The initial systems for generating deepfake content as shown in Fig 2, employed primarily autoencoder-type architectures, in which two encoders and decoders were trained jointly, with one encoder/decoder pair representing one identity and the other pair representing another identity, thereby enabling latent-space face swaps. Each encoder in these models learned embeddings inherent to a specific identity, and during inference the latent space of the encoder for one identity was provided to the decoder for the other identity, resulting in a realistic synthetic face that had undergone a swap of identity-related features. The classical two-stream autoencoder pipeline for facial manipulations fundamentally disrupted realism in deepfake content by reducing visible artifacts that early detection models relied upon. It is essential to understand those prior constructs to understand the backbone of current deepfake creation.

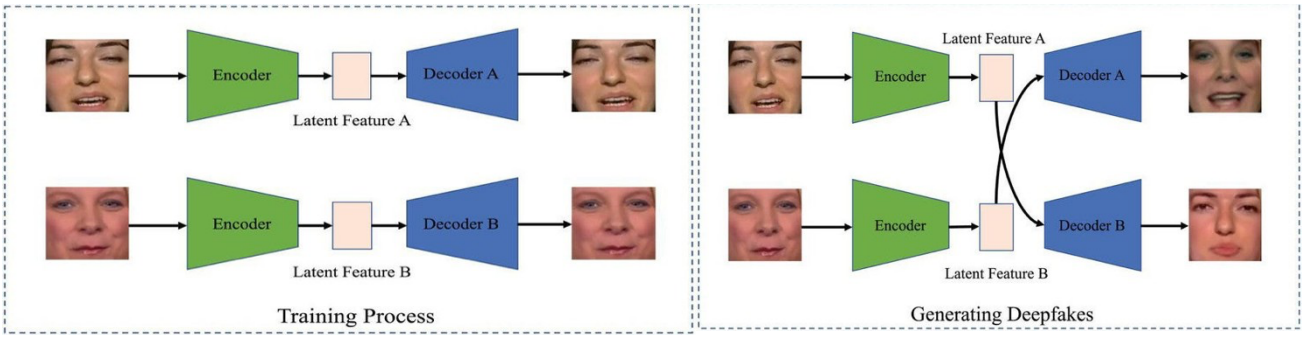

Figure 2: Traditional Autoencoder Deepfake Architecture

To address this gap, the present study proposes a hybrid GAN–CNN pipeline intended to increase the robustness of deepfake detection by utilizing a lightweight U-Net–based generative refiner, to contribute slight, adversarial perturbations to original frames or source frames to mirror the types of real-world distortions created in the face of compression, re-encoding, or post-processing. This should provide the detection model with a broader distribution of the manipulated content and may simply lead to greater detection capabilities in the case that manipulative generators evolve or improve even beyond current autoencoder structures. XceptionNet, chosen for

its strong performance in deepfake detection literature, is then fine-tuned on this augmented dataset to improve discrimination of subtle forgery cues.

Therefore, the main objectives of this research are to deal with the generalization limitations of CNN-based detectors trained on FF++, to measure the effects of adversarial GAN-based augmentation on classification robustness, to explore any performance improvements elicited by a hybrid GAN-CNN system; and to contribute a practical, computationally efficient training method, which would generalize outside of controlled datasets. This study aims to leverage adversarial refinement in combination with high-capacity CNN classification to help reconcile the gap between benchmark performance and real-world reliability providing meaningful advances to the domain of deepfake forensics.

2. Literature Review

Due to the rapid growth of generative models, deepfake detection has become one of the most pressing problems in the field of multimedia forensics. Generative adversarial networks (GANs), autoencoders, and diffusion models are highly advanced generative models that can produce synthetic media that is very difficult to separate from real content. Traditional forms of detection which rely on visible artifacts, and low-level pixel cues, are increasingly ineffective as generative models continually address the inconsistencies that early detectors relied upon. Therefore, there has been significant research into a variety of domains within computer vision, pattern recognition, and multimodal learning to develop detectors that can identify fine manipulations in a progressively real media rather than a usual quality of fake.

Paper [1] provides one of the most comprehensive systematic reviews of deepfake detection techniques, grouping detection technologies into three primary types: Convolutional neural networks, Recurrent neural networks, and Transformer-based technologies. The review discusses three critical areas: First, CNNs have a very high level of capability for detecting spatial artifacts within images and video; however, these technologies may have limitations when working with compressed images/video. Second, RNN technology can effectively model temporal relationships between visual frames; however, there are significant limitations regarding scalability. Third, transformer-based technologies may provide a possible avenue for further advancement but are currently expensive due to the high amount of computational resources required to operate such systems. One of the major issues that all future research efforts in deepfake detection will face is dataset overfitting. The paper notes that most of the existing deepfake detection systems work effectively with curated datasets, such as those available through FaceForensics++, but typically do not work well when applied to real-world data, which often contain noise, compression artifacts, and various manipulations. As a result, it is evident that enhancing generalization from training data (such as curated datasets) will be essential for developing effective real-time deepfake detection systems, and various augmentation strategies, including those developed in the present study, represent one possible means of addressing this issue.

Paper [2] presented the DF40 benchmark, which is designed to increase the scope and increase robustness and cross-domain variations of datasets. The DF40 contains 40 distinct kinds of manipulations, many of which were produced through modern diffusion models. The exploration completed by [2] indicated that even when state-of-the-art detectors have been trained with FaceForensics++ or DFDC datasets, their performance is greatly reduced when assessed on the DF40 benchmark. This finding reveals an important weakness within the current deepfake detection systems. Although the systems achieve nearly perfect performance when trained using traditional benchmarks, they are incapable of transferring to the increased variety of manipulative methods that exist within the broader space of actual online/social media. As such, this indicates that current systems are not representative.

of the types of manipulations that can occur within typical online/social media content. The conclusion of the findings of [2] support the use of augmentations to artificially diversify training distributions based on actual real-world distortions, which aligns directly with the GAN-based Refinement Pipeline from this study.

The misuse of deepfakes in societal situations, like political misinformation, identity theft and financial fraud, is the primary focus of the research outlined in paper [3]. Paper [3] argues that video detection systems need to be much more than just able to classify a video; they should also be constructed with the ability to withstand adversarial and deceptive attempts to bypass them. The need for video detection systems to withstand the common deformities encountered during the process of reposting, down sampling, re-encoding transformations commonly encountered on social networking platforms, is clearly shown in the research presented in [4]. Both paper [3] and the evidence presented in [4] confirm that it is equally important for classification accuracy and for resisting any real-world distortions.

The real-world robustness challenge is further described in paper [5], with the results of a comprehensive review of the autonomous detection systems, which indicates that virtually all detectors (CNNs, RNNs, and multimodal detectors) are vulnerable to such transformations as compression, noise, motion blur, and resolution shift. Furthermore, this review concluded that current methods of dataset generation do not adequately reflect real-world distortion thus, detectors will tend to overfit. Papers [6] and [7] Investigate lightweight detection architectures, with Paper [6] proposing a MobileNet-XG Boost hybrid, containing optimization for real-time application, and Paper [7] comparing VGG16, VGG19, and ResNet50 for image detection of forgery. Both Paper [6] and Paper [7] indicated that traditional CNNs do not create generalizable results when used outside of their existing training distributions. The results suggest that the benefits of improved architectural designs are only realized when using superior training data that has been made by a vast number of people in a wide array of locations/ethnic groups.

In paper [8], researchers propose a federated learning and blockchain-supported detection system with a focus on preserving privacy during collaborative learning. The concept for the project is novel, however, this system lacks resilience against the detection of previously unidentified types of manipulation. Another method presented in paper [9] uses fine-grained audiovisual inconsistency detection to find differences between a person's lips moving with their speech. This multimodal view provides good resistance to changes based solely on appearance; however, since it requires synchronized audio-visual inputs and good-quality video recording, its use will be limited in many venues where videos are published on the internet.

The paper [10] proposes a dual-branch model called LGDF, which incorporates global and fine-grained semantic information from images, providing good performance on benchmark datasets. However, Paper [10] finds that there are substantial losses of accuracy when applying this model to generalization across domains, which corresponds with the conclusions of Paper [11]. Paper [12] describes Faceswap Finder, a fusion-based method specifically developed for detecting face swaps. This technique does not perform well when applied to many newer neural texture and other forms of manipulated image technologies.

Research conducted in Paper [13], AdaBoost is evaluated using the DFDC dataset. The results show only limited improvements in the ability to detect simple manipulations, however the method is ineffective against higher quality synthetic images, as it relies more heavily on handcrafted features. Research in Paper [14] provides a comparison of CNN models and concludes that the Xception model yields the highest accuracy for detecting deepfake images among classical CNNs. However, this model was found to be very susceptible to adversarial noise, sparking the requirement for adversarially augmented training methods; one such method is GAN-based refinement, which has been utilized in the current research study. In Paper [15], there has been a shift away from

using CNNs, as ViT may provide superior analytical performance in terms of g eneralization across datasets when compared to CNNs, owing to their ability to identify long-term/long-range relationships between local/global features. Nonetheless, the high computational demands of ViTs makes the models unsuitable for low-power forensic applications.

Research conducted in Papers [16] and [18] pertains to the micro expression dynamic elements of accurate expression analysis, using an approach that combines both temporal modeling and masked autoencoders. The findings indicate that subtle facial signals can indicate an image may have been manipulated in addition to pixel-level inconsistencies but using cropped, high-definition facial images and properly aligned faces limits the effectiveness of these methods on detecting deepfakes from unsegmented video sources. A broad, in-depth review of deepfake creation and identification was provided in Paper [17], and it identified four significant concerns: reliance on datasets; inability to tolerate new forms of manipulation; no means for modeling how actual video degrading affects detection methods and inadequate investigation into how training enhancement from augmentation affects detector performance. These four issues create the need for techniques that enhance training datasets artificially. By doing so, the goal of making detection algorithms more resilient to deepfakes would be achieved.

The reviewed studies display one consistent theme. Detectors rely predominantly on the presence of artefacts from benchmark datasets when detecting deepfakes. As such, when real-world changes or manipulations of frames are applied to these artefacts, the detectors are unable to detect these changes. Previous work has concentrated primarily on the creation of new architecture (i.e., CNNs, RNNs, Transformers, Multimodal, Ensemble), while very few publications focus on the use of adversarial refinement of training data to improve generalization. This research fills the gap between prior research and the need for advancing detection methods using GAN-based U-Net refiner to create small perturbations that simulate real-world changes to the training data, thereby producing an 'adversarially improved' training dataset on which to fine-tune an XceptionNet based detector. Furthermore, as opposed to earlier research which improved model accuracy, the work proposed improves deepfake detector robustness and adaptability, and provides a new augmentation strategy, thereby bridging the gap between benchmark performance results in a controlled setting and the uncertain reality of real-world application in the field of deepfake forensics.

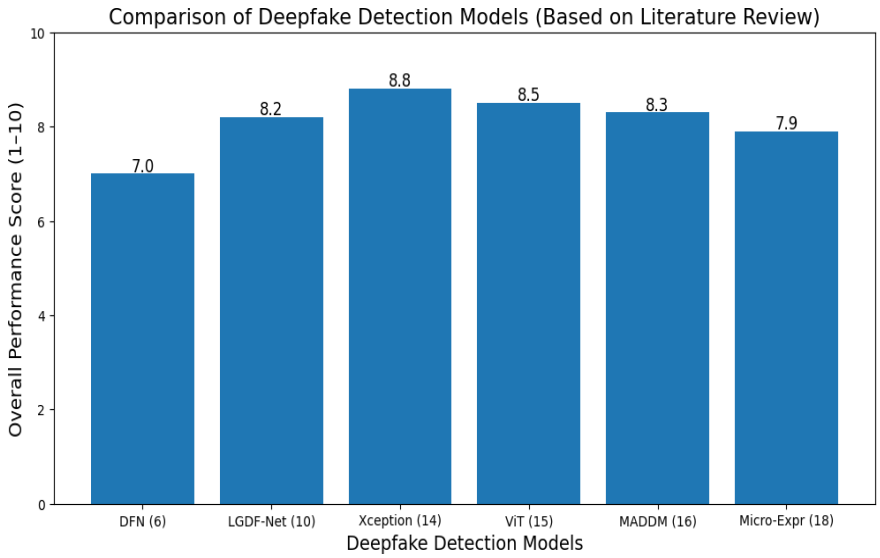

Figure 3: Comparative Performance of State-of-the-Art Deepfake Detection Models (Aggregated From Literature Review).

Table 1 : Summary of Literature Review

Ref	Tools Used	Advantages	Disadvantages	Performance
[1]	CNN, RNN, Vision Transformers (reviewed)	Broad analysis; identifies major detection weaknesses.	No model proposed; relies on surveyed work.	CNNs 85–98%, ViTs 84–90%.
[2]	DF40 Benchmark, SOTA detectors	Highly diverse benchmark, reveals robustness gaps	Very computationally heavy.	95% → 55–70% (cross-dataset).
[3]	Case studies, threat modeling	Strong real-world threat analysis.	No experimental detector proposed.	N/A
[4]	FF++, DFDC detectors	Shows real-world degradation under OSN distortions.	Still uses curated/test datasets.	Accuracy drops 15–40%.
[5]	CNN/RNN multimodal models	Covers image, video, audio deepfake detection.	Review only; no new architecture.	25–50% drop across domains.
[6]	MobileNet + XGBoost (DFN)	Lightweight & real-time friendly.	Weak against high-quality fakes	91–93% (FF++).
[7]	VGG16/19, ResNet50	Simple baselines; interpretable.	Outperformed by Xception/ViT.	~96% (VGG19).
[8]	Federated CNN, SegCaps, Blockchain	Privacy-preserving and secure.	Very expensive computation.	92–94%.
[9]	AV-sync CNN models	Audio-visual cue alignment improves robustness.	Needs clean audio + synced lips.	88–93%.
[10]	LGDF-Net dual-branch CNN	Strong on face-swap/identity forgeries.	Fails under domain shift.	96% → 70% cross-dataset
[11]	CNN, RNN, ViT	Holistic architectural comparison.	No new model proposed.	85–99% → <70% cross-domain.
[12]	CNN fusion	Effective for face-swap detection.	Weak on neural-texture fakes.	~94%.
[13]	AdaBoost	Simple and interpretable.	Poor against GAN/diffusion fakes.	82–87% (DFDC).
[14]	Xception, ResNet, VGG	Xception validated as strong baseline.	Very adversarial-sensitive	Xception: 98.3% (FF++).
[15]	Vision Transformers (ViT)	Best cross-dataset generalization.	High computational cost	84–90% cross-dataset.
[16]	Masked Autoencoder (MADDM)	Captures subtle micro-expression cues.	Requires high-quality aligned faces.	92–95%.
[17]	GANs, Autoencoders, Transformers	Identifies open challenges & future gaps.	Survey-based; no experiments.	Typical 85–99% ranges.
[18]	CNN + Temporal Micro-expr. Model	Strong against neural-texture forgeries.	Fails on low-quality/noisy videos.	90–94%.

3. Findings and Related Work

The existing literature review on deepfake detection reveals some similar trends when comparing traditional CNN-based approaches to transformer models, multimodal frameworks, and ensemble techniques. First, nearly every detection system reported strong performance on controlled academic benchmarks like FaceForensics++, Celeb-DF, and DFDC, all of which created manipulated files using a complete generation pipeline that was defined within the test datasets. Most of the resulting accuracy values exceeded very high (90%–98%) levels for many architectures (e.g., CNN's like XceptionNet, LGDF-Net, VGG's) that have been evaluated for these types of tests. However, it is rare for these results to transfer well into the real world. As demonstrated in both DF40 and cross-dataset evaluations, the performance of the model tends to decrease dramatically (-30% to -50%) when testing on files created outside of the original test environment (e.g., unseen manipulations, compression artifacts, videos with low-light levels, and forgeries produced by diffusion models). This demonstrates a key vulnerability in the current models, they associate patterns that are unique to datasets as opposed to developing generalizable manipulation-invariant features.

The second major finding across comparative studies is that the robustness of the model relies heavily on the number and quality of different training datasets. Traditional models (such as DeepFakes and Face2Face) create different error patterns than modern models that use modern diffusion approaches (Neural-texture models) and the most advanced GANs (like StyleGAN3). Because of this, most detection models have been trained on relatively small numbers and do not capture the complete range of manipulative techniques and will overfit to particular patterns specific to the datasets used in their development. In the literature study, CNNs learn low level pixel inconsistencies better than Vision Transformers (VITs) and VITs outperform CNNs at learning and generalizing from more complex, globalized semantic information. But both classes of models do not offer sufficient robustness to real-world causes of variation, for example, re-encoding noise, jitter between frames, stabilization artefacts, and resolution changes show a continued gap between what can be achieved in the lab and what is achieved when distributing the model across varied data environments.

Third, while multimodal detection has become more advanced using speech-lip dynamic alignment, micro-expression tracking, and ensemble fusion as shown in Figure 4, it continues to rely heavily on having high-quality input. If the audio is incorrect, the face is blocked, or there is inconsistency in the movement, multimodal detectors cannot perform as intended. Surveys and benchmarking studies also highlight a second major vulnerability: susceptibility to adversarial manipulation. A small change to a pixel could cause them to incorrectly categorize their detection, which highlighted the need for training methods that can withstand adversarial attacks.

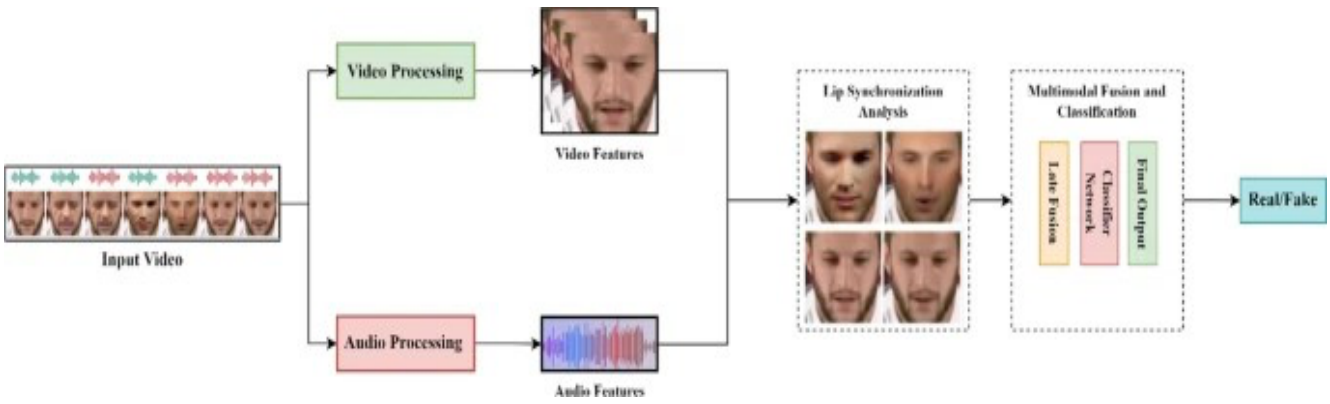

Figure 4: Example of modern multimodal deepfake detection pipelines that integrate video features, audio features, and lip-synchronization analysis before classification.

Additionally, there are no existing literature that focus on explicitly adding synthetic or adversarially perturbed images to existing training data to emulate the type of noise and distortions encountered by deepfakes in a real-world setting. Most of the research have concentrated on model architectures and feature extraction, whereas very few have focused on the specific process of adding adversarial refiners, perturbative GANs, and noise-conditioned generators to the training set. Considering that real-world versions of deepfake content will pass through numerous platform related processes, including loss during recompression, altering domains, and developing artifacts, this creates a substantial gap in the current research.

3.1 Research Gap Addressed by this Work.

Based on these findings, the main shortcoming identified is the lack of an effective strategy for detection quality through a focus on data-related elements. The primary reason current detection methods do not perform well is not since they were unable to identify forgeries; Rather, these systems are trained using highly curated clean datasets, which do not reflect the noise, distortion, or perturbations present in typical video files from the field.

The proposed research addresses this gap by introducing a lightweight U-Net based GAN that generates adversarially refined fake samples that were derived from FF ++ videos. The perturbations are meant to reproduce degradation and distortion in the real world, such as blurring, color shifts, frequency distortions, and smoothing of artefacts. With the augmentation of XceptionNet's training dataset, the addition of these samples will expose the model to multiple types of forgery patterns, allowing it to learn features that are invariant to the manipulation process used, rather than features that are based on the training dataset alone.

This approach directly fills the gap identified in the literature by:

- Implementing a data-centric robust layer, which is missing within the existing architectures.
- Implementing real-world compression and noise effects during training, rather than during testing.
- Improving the generalization capability without the need for multimodal data, which is required by many of the detection tools.
- Maintaining computational efficiency, as opposed to other (ViT) or multimodal approaches to detection.
- Reducing adversarial vulnerability, which has been reported in numerous previous studies.

Therefore, the hybrid GAN-CNN pipeline helps to bridge the gap between the accuracy measured in benchmarks, and the real-world reliability of test results; and it provides a realistic and scalable solution for enhancing deepfake forensics.

4. Dataset Description

4.1 FaceForensics++ Dataset

The FaceForensics++ dataset (figure 5) is the main benchmark for this study due to the diverse range of manipulation methods and compression rates as well as the number of different identities represented in the data collection. The 1,000 pristine videos that comprise the FF++ dataset was obtained from YouTube and have a diverse population of ages, ethnicities, facial features, and environments. These videos represent a solid basis to generate realistic deepfake manipulations. The FF++ dataset also contains a further 4,000 manipulated videos that were created using the four approaches of Classical Techniques (CT), DeepFakes, Face2Face, FaceSwap, and Neural Textures; thus, providing data to evaluate both Identity Swap and Expression Transfer forgery manipulation methods. As such, the FF++ provides an excellent testing ground for assessing the robustness of models against a wide range of forgery pipelines.

FF++ provides a consistent form of annotation across dataset resources, a standardized video format, and processed bounding boxes for each face within a video resource, which provide the same level of confidence in test methods and reduce the likelihood of discrepancies between tests conducted using different methods. In addition, the fact that FF++ is widely used by researchers allows for the ability to compare multiple methods on the same dataset using standard metrics. Furthermore, its balanced composition of real and fake videos provides equal representation of each type of video resource to avoid bias in the sampling of binary classification tasks (e.g., real, or fake). All these elements together make FF++ the best choice for developing strong and generalizable deepfake detection systems.

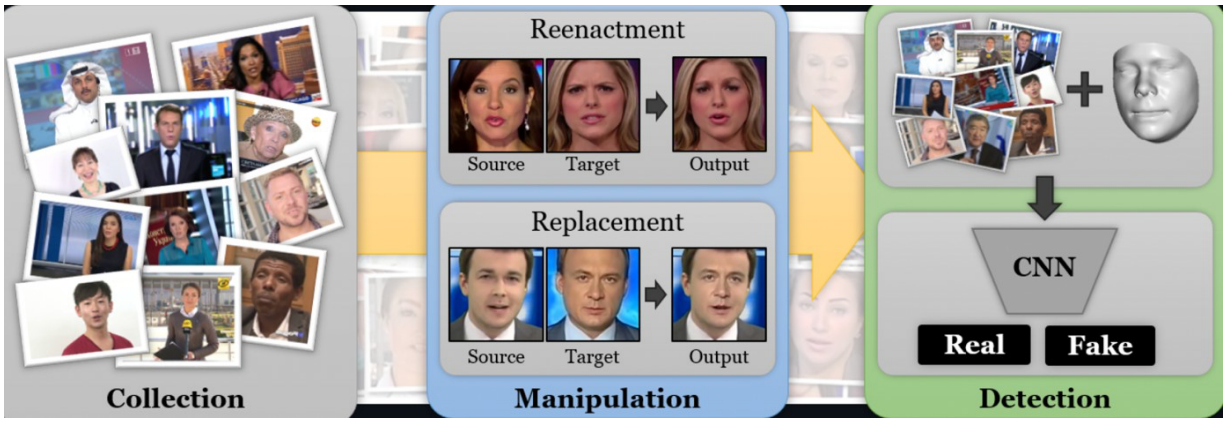

Figure 5: FaceForensics ++ dataset

4.2 Frame Extraction and Face Cropping

The videos obtained from FF++ and used as training samples were each broken down into separate frames shown in figure 6 using a uniform time step across the entire set of videos. Rather than taking all frames, which would result in a considerable amount of duplicate information in time, only every k'th (or for example, every 10^{th}) frame was extracted. This way not only enough of the range of time is covered by each emotion/pose, but it also limits the total number of images and avoid introducing any bias to the training by having multiple almost identical crops from multiple images. Each extracted frame is saved with a unique identifier based on the name of the video it came from, allowing for matching between frames and the video they were taken from at later stages of the training process.

A face detector using MTCNN is applied for every frame. If there is at least one face detected, the largest bounding box will be selected as the bounding box of the face of the primary speaker. The detected bounding box will then be cropped with about 10% to 15% margins from the left and bottom edges and will be padded with zeros to match the standard input shape of the GAN and XceptionNet models (usually 256*256 pixels). All faces have been resized to have the same aspect ratio and pixel dimensions, providing consistency for stable training of the GAN refiner and XceptionNet detector. All frames where no faces are detected are not included in the training data so that samples that do not contain any human faces or are background will not be included in the training process.

MTCNN cropping provides a concentrated learning signal on the facial region where manipulations primarily happen, as opposed to diluting the model with non-related background pixels. Additionally, it indirectly normalizes the position and scale of the head which reduces the potential for variability to be incorporated into the model in non-useful ways. By combining face-centric cropping with temporal subsampling, the resulting set

of face images from the preprocessing pipeline can be placed into a high-density but non-redundant alignment that is effective for both convolutional architectural methods and for generative refinement processes.

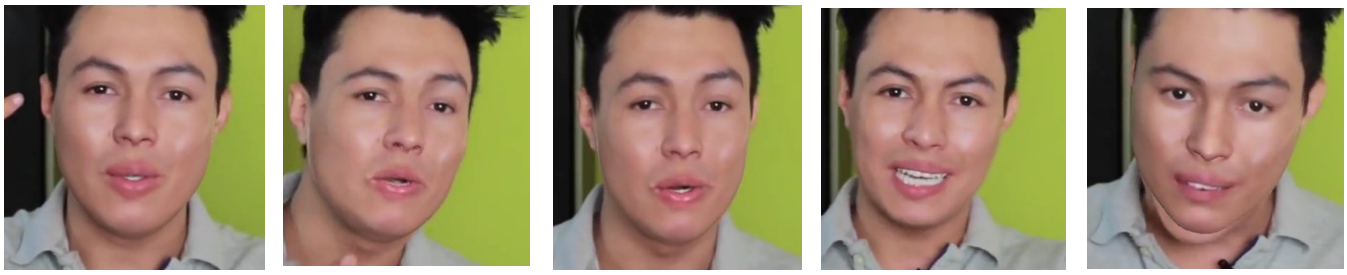

Figure 6: Sample cropped frames of a deepfake video using MTCNN

4.3 Dataset Organization and Class Balancing

Once the face images are cropped, the next step is to assign a binary label to each image, depending on the source video from which it was created. A binary label is assigned to each cropped image as either "real" or "fake". If the cropped image comes from an untainted FF++ video, the binary label assigned to it will be "real", and if it came from manipulated versions (DeepFakes, Face2Face, FaceSwap, NeuralTextures), then the binary label will be assigned as "fake." The cropped images are structured in a regular manner by creating a classification directory tree under the top-level directory "reduced_dataset," which contains the following subdirectory structure: train/real, train/fake, val/real, val/fake, test/real, and test/fake. This structure will allow for better compatibility with standard PyTorch and TensorFlow dataset loaders, allowing for easier testing and subsequent use of the dataset.

Splitting the videos maintains a strong correlation between neighboring frames (data sets) preserving confidentiality, etc. with most videos allocated to the training dataset (70%); validation (15%); and test (15%). Thus, ensuring a variety of identities and manipulation types within each split, allowing a realistic evaluation environment with a sizeable training dataset suitable for deep learning.

FF++ may appear to have an equal number of genuine and counterfeit clips. However, at the frame level, there may be some disparities in whether a clip is correctly classified as fake or real due to differences in detection success rates and how frames are temporally sampled; thus, resulting in a potential bias towards the dominant class when training on such data. To resolve this issue with bias, three basic balancing techniques are applied when assembling the datasets: under sample the dominant category, oversample the less populated category, and use a class balanced sampler that performs best with the distribution of classes within the data loader. The validation/test datasets are to maintain a closer resemblance of the distribution of the natural classes, to provide an unbiased estimate of actual performance on imbalanced datasets found in everyday scenarios.

5. Methodology

5.1 Overview of Hybrid GAN–CNN Pipeline

The proposed method figure 7 employs a hybrid pipeline that combines a generative model with a discriminative detector. The primary goal is for the lightweight U-Net refiner to create hard examples by generating subtle perturbations of real faces that still look visually plausible but make it increasingly difficult for the detector to differentiate between the original frame and its variant. By creating these hard examples, the U-Net refiner

encourages the classifier to learn robust, manipulation-invariant features. Instead of creating new identities with a generator, the goal of the generator is to act as a residual refiner to both the real and fake frames of FF++. The generator introduces realistic distortions to both types of frames that typically occur as a result of compression/re-encoding and post-processing.

The training of the proposed method proceeds in three phases. The first phase involves training an XceptionNet detector on the original cropped frames of the downsampled FF++ dataset; this phase creates a strong content-only baseline for training the GAN in the second phase and allows the creation of a fixed discriminator in the GAN framework. The second phase consists of training the U-Net refiner, which creates a small residual for every input face while keeping the detector fixed. The residuals created by the U-Net refiner will direct the predictions of the detector to confusion (e.g., making the output of the detector for real faces more closely resemble that for fake faces and vice versa) while also adhering to perceptual and regularization constraints. The adversarial nature of this training exposes the U-Net refiner to the classifier's decision boundary and teaches the U-Net how to modify real faces to maximize confusion for the detector.

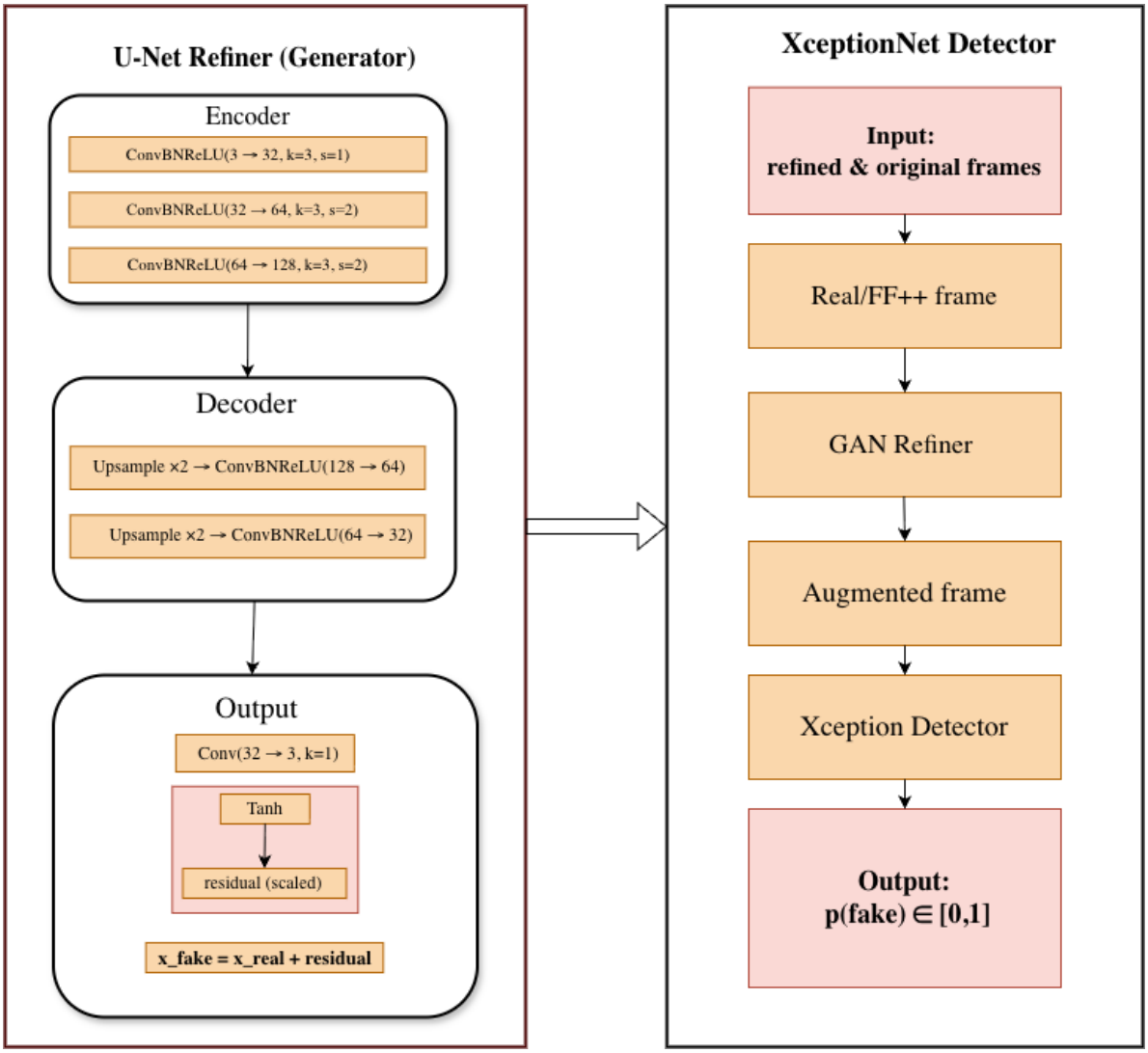

Figure 7: Proposed hybrid GAN–CNN deepfake detection pipeline

In the final stage, a combined dataset was formed by processing a group of the original training images through the trained generator and saving the generated refined images. The dataset was then used to fine-tune XceptionNet, the detector. During fine-tuning, a reduced learning rate was usually used, and the earlier layers of the network were partially frozen to preserve the previously learned representations. A dataset is created that has been built from images taken from both the original training dataset and from those produced by using Generative Adversarial Networks (GANs), thereby providing both types of images to XceptionNet. This has enabled XceptionNet to learn how to detect the differences more effectively, as well as prevent it from being over-trained by any one way of producing images, which can introduce artefacts.

5.2 U-Net Refiner (Generator)

The generator G is a subtype of a U-Net that is compactly built as a Refiner, which will accept an image x of a normalized face x∈[0,1]3×H×W and output a small residual image r that has the same dimensions as x. The encoder structure of G is composed of three ConvBNReLU blocks. The first ConvBNReLU block maps three channels of input into 32 channels, at the full resolution of the input image. The subsequent ConvBNReLU blocks will down-sample the original images by two times (stride s=2) and increase the number of channels to 64 and 128 as well. This hierarchical organization of the encoder allows for an increased global understanding of the structure of a face, while maintaining a relatively low number of overall parameters within the model.

The decoder has two stages of upsampling and is a mirror image of the encoder. The feature maps produced in this step are upsampled to the original resolution with bilinear interpolation (scale factor 2) and fed through a series of ConvBNReLU blocks to reduce the number of channels from 128 to 64 to 32. In order to maintain the fine grain structure of images, skip connections from the encoder to their corresponding decoder stages can be added, though for simplicity and stability this implementation does not include them. At the last stage of the decoder, a 1x1 convolutional layer reduces the number of channels from 32 to 3, followed by a tanh activation layer that scales the result by a small constant (for example, 0.05) to generate the residual r. The refined image can be computed as x′=clip(x+r) to ensure that the pixel intensity values are valid.

To train G, its loss function is made up of multiple components. The adversarial loss allows refined images to change the frozen detector's prediction to be closer to a specified target which may include increasing the fake probability of a real image by the detector or having fake and real images closer to the decision boundary of the detector. To constrain the size of the residuals between x and x' so that the perturbations are subtle, visually imperceptible, the L1 reconstruction loss encourages the learned reconstructions to remain close to the original input images. In addition to these two components, it is often helpful to impose a frequency-domain penalty (e.g., high frequency of r) on the generator to prevent it from inserting noise that appears unnatural into the image refinements generated by the generator. Collectively, these components help ensure that the generator learns to make realistic, detector-aware refinements of the images as shown in figure 8 instead of creating obvious artefacts.

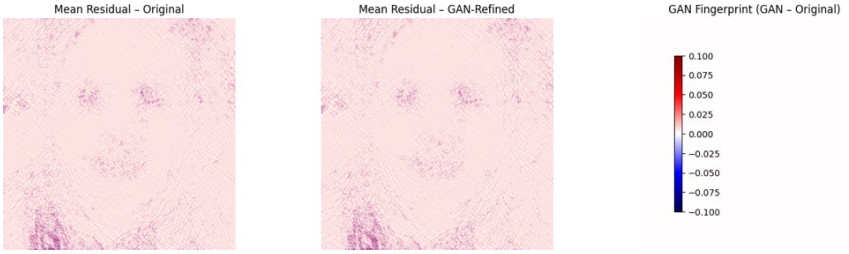

Figure 8: GAN fingerprint extraction using SRM high-pass residual averaging.

Table 2 : U-Net Refiner Generator Architecture

Layer	Output Shape	Details
Input	3×256×256	RGB face crop
Down1	32×256×256	ConvBNReLU(3→32, k=3, s=1)
Down2	64×128×128	ConvBNReLU(32→64, k=3, s=2)
Down3	128×64×64	ConvBNReLU(64→128, k=3, s=2)
Up2	64×128×128	Upsample×2 + ConvBNReLU(128→64)
Up1	32×256×256	Upsample×2 + ConvBNReLU(64→32)
Output	3×256×256	Conv(32→3, k=1), Tanh, scaled residual

5.2 XceptionNet Detector and Training Strategy

The design of the detector D employs the XceptionNet structure, as it has provided the best results in prior research regarding deepfake detection. The network's weights are first established on the ImageNet dataset, it has already been exposed to the largest collection of image data available. The output for each input frame has been set to return a range from zero to one based upon the use of a one-neuron final classification layer and the use of a sigmoid activation function. The detector will output $p\_fake \in [0,1]$. Input images will be pre-processed to have their required sizes (such as 299×299) as shown in algorithm 1 , be normalized using ImageNet mean and standard deviation values, and have been augmented during training by randomly flipping images horizontally and adjusting contrast and brightness levels.

During the baseline training period, the XceptionNet model is trained only using binary cross-entropy loss with an Adam optimizer with the original FF++ crop data. To balance the classes, the sample based is either weighed on how frequently they appear in our data set or the losses in back-propagation process are weighed so the advantage isn't given to a particular class due to the number of samples in that class. In this phase, the model was trained for 25 epochs , and tested at various points by using early stopping and reducing learning rates if no improvement is noticed in ROC-AUC and F1 scores during validation phase. After completing the baseline training phase, the detector performs well against the FF++ in-distribution data. However, it may be susceptible to changes between the domains as well as to subtle manipulations of images that were not seen during training.

In the GAN augmented fine tuning phase, a new training dataset was created that contained images with and without enhancement by the GANs .For each mini-batch 50% of the contents were images without enhancement and the other half were enhanced to allow the detector to consistently see both types of the image in training. During this fine tuning process the first layer of XceptionNet were frozen to maintain their ability to capture generalised low-level features while the more complex layers were refined to extract features from difficult samples that had been created by the GAN. Finally, the fine tuning process was conducted at a low learning rate over a shorter duration and was monitored using validation metrics, after completing the fine tuning process the best performing model was evalutated on held out test data set along with external images/videos where the proposed hybrid GAN-Xception pipeline achieved an accuracy of 95.1% with 97.2% precision, 93.3% recall, 95.2% F1 and 98.5% ROC-AUC.

Algorithm 1: Data Pre-Processing for XceptionNet Detector D

Input: Reduced dataset (train/, val/, test/), GAN-refined images

Step 1: Load the reduced FF++ dataset from the given directory.

Step 2: Apply image cleaning: remove corrupted frames, resize all images to 256×256 and center-crop to 299×299.

Step 3: Perform Data augmentation (training only)

- Random horizontal flip,
- Light color jitter,
- Convert to tensor.

Step 4: Normalize images using ImageNet mean and standard deviation.

Step 5: Create balanced batches with equal numbers of real and fake samples.

GAN-Augmented Fine-tuning Phase

Step 6: For each image x, generate a GAN-refined version $x' = G$ and save into train_aug/real or train_aug/fake.

Step 7: Construct a mixed dataset (50% original + 50% GAN-refined).

Step 8: Apply the same pre-processing pipeline (Steps 2–4).

Step 9: Fine-tune XceptionNet using a reduced learning rate while monitoring validation accuracy, F1-score, and AUC.

Function LoadDataset(path):
dataset ← read all images from path
Return dataset

Function CleanAndResize(image):
image ← remove invalid/corrupted data
image ← resize(256×256)
image ← center_crop(299×299)
Return image

Function Augment(image):
aug_image ← random_flip(image)
aug_image ← color_jitter(aug_image)
Return aug_image

Function Normalize(image):
norm_image ← (image − μ_ImageNet) / σ_ImageNet
Return norm_image

Function GANRefine(image, G):
refined ← G(image)
return refined

Function BatchCreator(original_set, refined_set):
batch ← sample 50% from original_set and 50% from refined_set
return batch

Output: Pre-processed and augmented batches for XceptionNet training and evaluation

5.3 Design Rationale

A hybrid GAN-CNN system was developed to overcome the two main problems identified in the literature: (i) CNN-based methods have poor generalization ability for new forms of manipulation; and (ii) modern deepfake artefacts are becoming increasingly subtle as technology improves. The U-Net model is most effective as a generator due to its ability to provide enhanced spatial coherence during image enhancement while maintaining semantic integrity of the facial identity. Compared to current state-of-the-art GAN models, such as StyleGAN and DeepFake autoencoders whose large sizes result in several unintentional identity and geometrical changes, the U-Net model is capable of producing only small noise such that the overall altered facial features of the produced image remain close to its original version.

The XceptionNet architecture was selected as the main detection method because it incorporates depthwise separable convolutions, which have been empirically demonstrated to perform well for detecting deepfakes on multiple datasets. In addition to having a larger receptive field than the established alternative, which allows for more extensive extraction of texture and frequency-aware patterns, XceptionNet's ImageNet-trained weights provide a good source of reliable low-level features. Combining both generator and detector will allow the generator to create more difficult-to-detect refined samples while the detector learns manipulation-invariant features that improve over time.

5.3.1 Loss Formulation

The generator G is trained using a composite loss function:

$$\mathcal{L}_G = \lambda_{adv}\,\mathcal{L}_{adv} + \lambda_1 \parallel x - x' + \lambda_{freq}\,\mathcal{L}_{freq} \quad (1)$$

- Adversarial Loss:

  $$\mathcal{L}_{adv} = -D(x') \quad (2)$$

  forces refined images to push detector predictions toward confusion.

- L1 Reconstruction Loss:
  Constrains the generator so that residuals remain subtle and visually imperceptible.

- Frequency Regularization:

  $$\mathcal{L}_{freq} = \parallel FFT(r) \parallel_1 \quad (3)$$

  penalizes unnatural high-frequency patterns.

This combination ensures that refinements remain realistic while still impacting the detector's decision boundary.

5.4 Training Hyperparameter

Table 3: Training Hyperparameters

Component	Optimizer	LR	Batch Size	Epochs
Xception Baseline	Adam	1e-4	32	25

U-Net Refiner	Adam	2e-4	16	20
Xception Fine-Tuning	Adam	5e-5	16	10

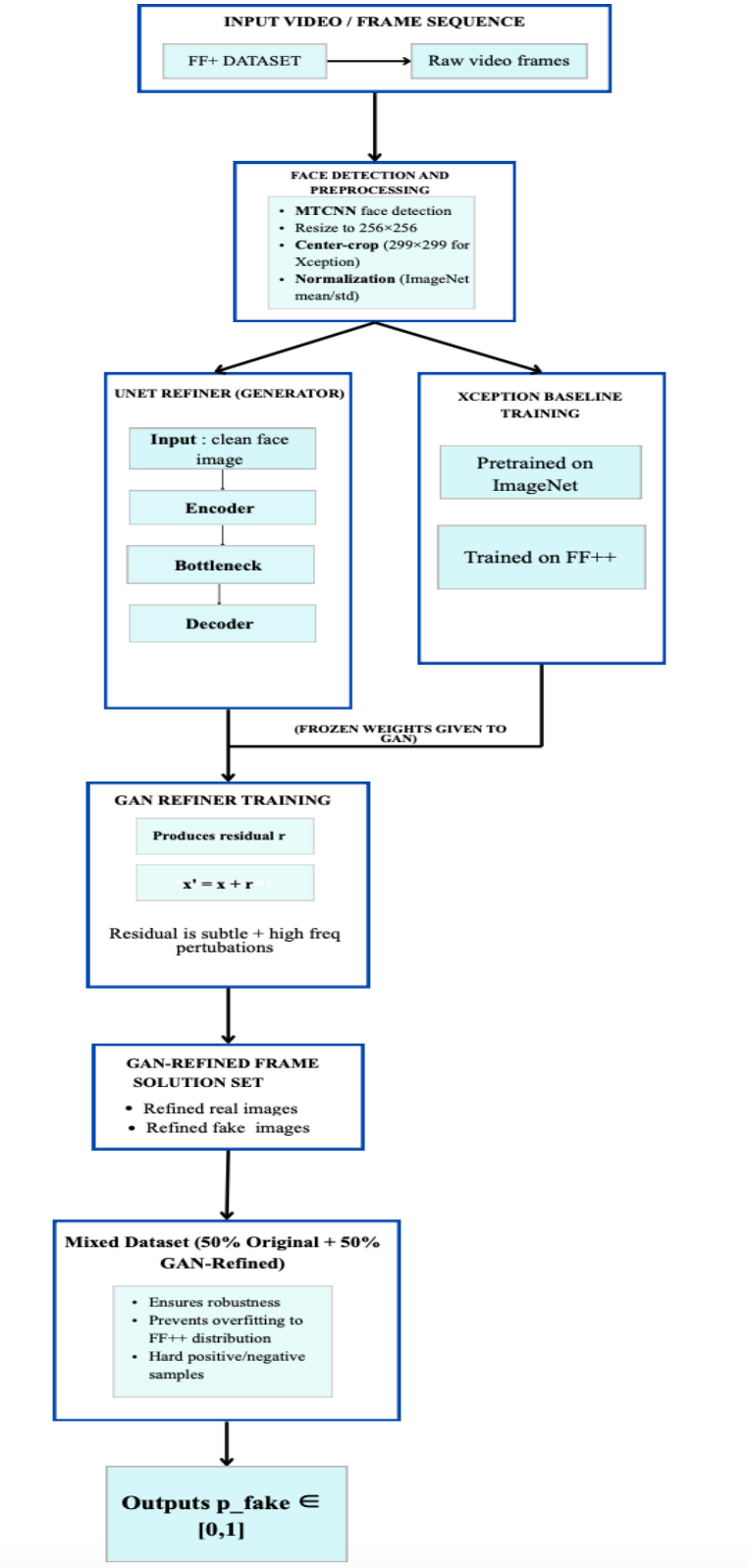

Figure 9: End-to-end pipeline summary of the proposed methodology

5.4 Model Evaluation Metrics

The performance of the suggested deepfake detection models was quantified using normalized classification metrics derived from confusion matrices. The accuracy is computed as the total number of correctly identified samples divided by the total number of sampled objects.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (4)$$

To evaluate per-class behavior, the Precision and Recall metrics were calculated. Precision reflects the proportion of fake predictions of all the fake samples correctly classified by the model. Precision can be defined mathematically as:

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

While Recall is used to measure how well a model has identified all the fake instances. The mathematical expression for Recall can be represented as:

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \quad (6)$$

The F1-score combines the two metrics together to measure overall classification performance via the harmonic mean. The F1-score is calculated as:

$$\text{F1 } Score = 2\,X\,\frac{Precision\,x\,Recall}{Precision+Recall} \quad (7)$$

The ROC/AUC coefficients are independent of the threshold and measure how well separate the real class and fake class is. An ROC curve is a graphical representation of the True Positive Rate (TPR) vs. False Positive Rate (FPR).

$$TPR = \frac{TP}{TP+FN} \quad (8) \qquad FPR = \frac{FP}{FP+TN} \quad (9)$$

The area under the ROC curve (AUC) summarizes the ROC curve to provide a single number quantifying the model performance of the classification system. All these evaluation metrics will provide a complete and thorough evaluation of both the baseline XceptionNet and the GAN-augmented detectors.

6. Results

6.1 Baseline XceptionNet Performance on FF++ Dataset

When training the baseline detector using only the downsized FF++ dataset, the baseline detector exhibited exceptional ability to correctly identify the real and false mark from any given sample. The XceptionNet model consistently improved in its training/validation accuracy at all four different eco-reduction levels through its time-matching of losses over thirty epochs; this shows good convergence as well. The XceptionNet model's training

using both standard augmentations (i.e., flip and color transform) and the ImageNet pretraining set also helped to stabilize the model's accuracy.

The findings of a quantitative assessment of the baseline classifier presented: accuracy = 90.91%, precision = 91.5%, recall = 91.3%, f1 Score = 91.0% & ROC-AUC = 91.0%. The classification performance of the built deepfake detection Classifier using the FF++ training data demonstrated validity within the FF++ Model Tests controlled environment, as the ROC displayed as shown in figure 10 displayed a high coefficient of increase toward the upper left corner of the graph and indicated that the baseline Classifier has the capacity to distinguish between the two class types with very little uncertainty about the threshold used to classify the images. The results of the confusion matrix as shown in figure 11 demonstrated a correct classification of many of the fake or manipulated frames by the detector; however, some were incorrectly classified when using highly compressed frames as real or authentic, this finding aligns with the overall trend associated with FF++ Detectors trained on FF++ data.

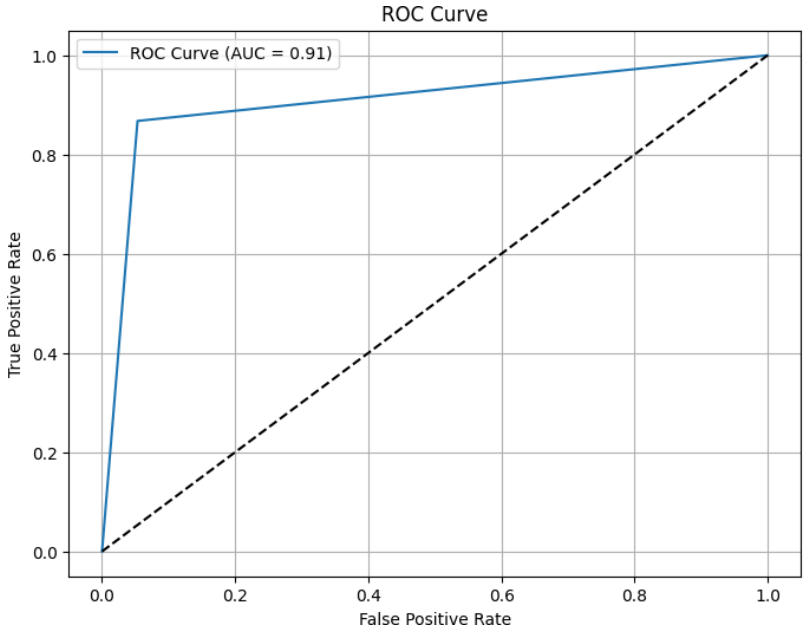

Figure 10: ROC Curve of the baseline classifier

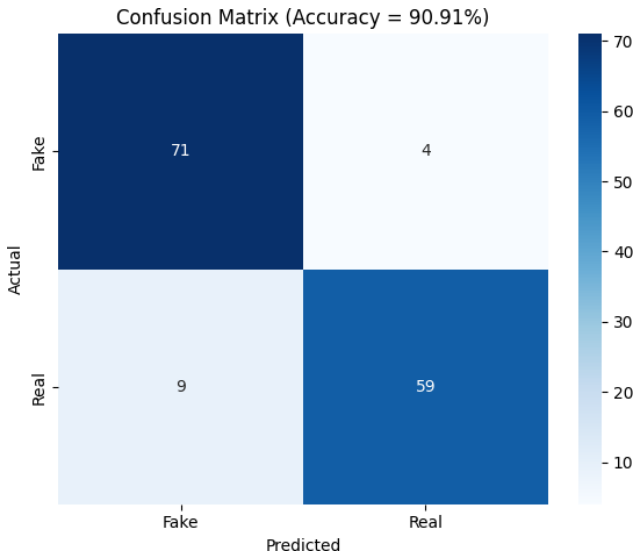

Figure 11: Confusion Matrix of the baseline classifier

The baseline performance for the trained models showed limited transfer quality when comparing results on previously unseen test samples of the data set referred to as FF++. On the other hand, internet deepfakes may exhibit a variety of subtle generative artefacts that are not present within the training distribution. Furthermore, test samples containing frames that exhibit lighting variations, were subject to re-encoding compression or were printed from different domains displayed a significant degradation in the performance of the model. Given these observations, this research pursued the integration of a GAN-based refiner to allow for an extension of the training distribution, thereby increasing the ability for the model to detect subtle cues related to manipulation.

6.2 GAN Refiner Outputs and Adversarial Perturbation Behavior

The stable performance exhibited by the lightweight U-Net based generator during training produced perturbations that were imperceptible to humans and nevertheless highly effective as adversarial examples. The GAN learned to manipulate the images such that they perturbed the confidence boundaries of the XceptionNet, while at the same time maintaining a coherent structure. Upon visual inspection, the GAN-processed images still appeared as they did before processing, but with high-frequency perturbations coming from the facial areas where manipulative cues exist.

The generator's operation was further revealed through residual maps, FFT spectra and GAN fingerprints as shown before in figure 8. In using residual heat maps, subtle distributions of energy were observed in the eye, lip, cheek and skin texture areas where many deepfake artifacts occur. In the frequency domain, GAN refined frames had higher frequencies relative to other methods, likely due to the generator providing additional detail in a manner consistent with the noised patterns of real-world noise associated with the compression and re-encoding processes. By artificially creating such noise patterns in GAN-generated images shown in figure 12, these modifications made classification more difficult, resulting in "hard positives" and "hard negatives" for use in training.

By diversifying the training set with the aid of the GAN refined training set, the trained detector became capable of generalising to samples that had never been previously shown to it. The generator was able to produce realistic distortions that introduced an increase in the number of ways to manipulate the image while maintaining global semantics. It was this capability which allowed XceptionNet to build an increased number of manipulation-invariant representations during the fine-tuning phase.

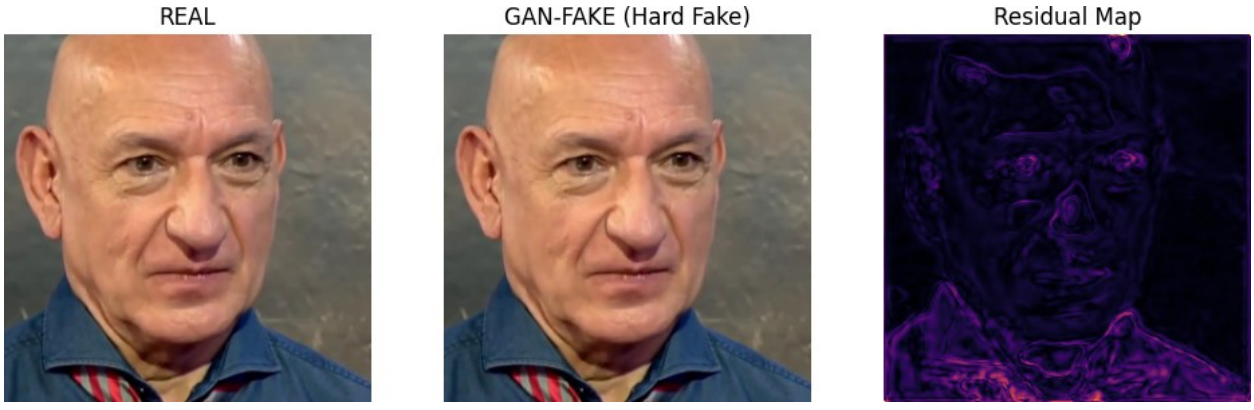

Figure 12: Real vs Gan-Fake: Subtle perturbations were introduced which is imperceptible to human eyes.

6.3 Mixed Dataset Fine-Tuning and Final Model Performance

Improving the performance of XceptionNet through fine-tuning (FF++) on a mixed dataset made up of 50% original and 50% GAN-refined images led to large improvements in every metric evaluated. This is because the GAN-refined (RFF) images served as a regularizing factor for the model, allowing it to create a larger and broader learned feature space (training image) while not overfitting FF++ specific patterns. The freezing of the early layers preserved low-level features that were generalizable across images, while the model could still adjust to more complex perturbations (i.e., noise) created by GANs in deeper layers.

The newly created hybrid generator-adversarial network (GAN) and the convolutional neural network-based Xception model obtained 95.1 % accuracy, 97.2 % precision, 93.3 % recall, F1 score of 95.2 %, and ROC/AUC area of under the curve (AUC = 98.5 %). This new hybrid GAN Xception model has improved the baseline model from all four perspectives.

A crucial outcome of the study was the ability to apply the hybrid models' general theories of learning using most major public datasets (FF++, etc.), No inputs used to train any of the neural net models were from any of these other public datasets. As a result, the hybrid model could accurately identify diverse manipulated media shown in figure 14 across different domains without suffering from detection errors or increased confusion with respect to the baseline detector. This outcome supports the need for further investigation into the utility and performance of adversarial augmentation techniques for improving out-of-distribution robustness.

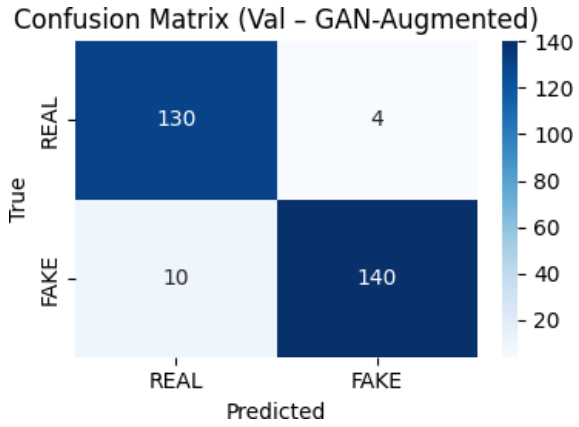

Figure 13: Confusion Matrix of the final Gan- Augmented Xception net model

Figure 14: Frame wise deepfake video inspection sourced from YouTube video.

6.4 Model Explainability and Grad-CAM Interpretation

Grad-CAM visuals were created for the baseline XceptionNet detector and GAN-augmented detector to gain more insight into what the models learned. The baseline model had strong attention on sharp edges and on mouths as well as on differing colors which comes from being a common FF++ artefact zone; however, little attention was paid to more discreet types of manipulations which caused the baseline to achieve low accuracy rates when there were no indications of an artefact, or when artefacts were hidden behind compression in deep fake pictures used in real life.

The use of Grad-CAM with the baseline XceptionNet mostly shows vague areas on the middle of the face shown in figure 15 since it is dependent on the features of the data set it was trained on (FF++), specifically due to the texture artifacts and blending discrepancies. Therefore, it produces large blobs of heat on the nose and mouth, which are the areas where current deepfake methods introduce characteristic distortions. On the other hand, the refined detector based on GAN produces a more localized activation with a higher degree of specificity to the local area and contains more of the smallest (micro) details shown in figure 16. The focus is on the edges of the image and the boundary between the face and the background; the refined detector is more focused on the local part of the facial structures. Because the U-Net model generates very fine and very small local perturbations during training, the refined detector becomes less reliant on artifacts on the surface and learns and utilizes the subtle high-frequency cues instead. For this reason, this GAN-based model has a greater degree of generalization and robustness by focusing on well-defined intrinsic inconsistencies across the different methods of generating deepfakes and a variety of compression rates.

The results of the above explained the explainability hypothesis that GAN Refined training samples provided the detector with a broader generalization of meaningful features, through the use of adversarial augmentation, will enable the hybrid model to perform better on unseen deepfakes.

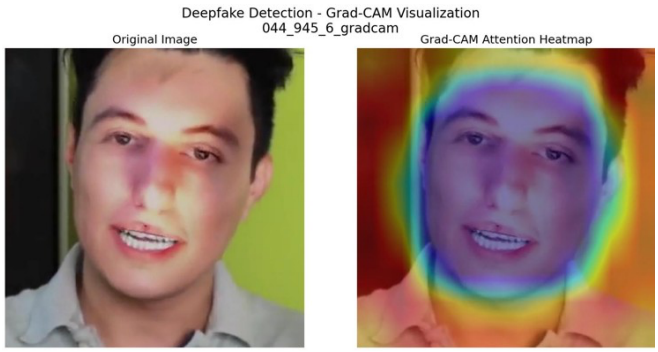

Figure 15: Gradcam Explanation of Baseline Xception net classifier

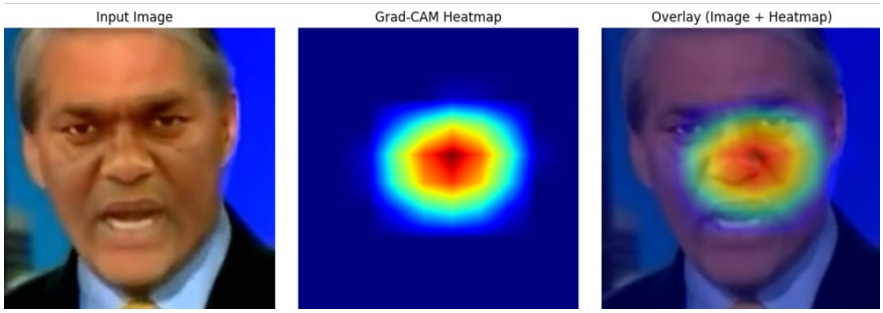

Figure 16: Gradcam Explanation of GAN augmented Xceptionnet classifier

Table 4: Classification Performance of Baseline vs. GAN-Refined XceptionNet

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline Xception Net	90.91%	91.5%	91.3%	91.0%	91%
GAN- Refined Xception Net Model	95.1%	97.2%	93.3%	95.2%	98.5%

7. Conclusion and Future Scope

This work describes a Deepfake Detection Framework that combines the features of both U-Net based GAN (Generative Adversarial Network) refiners and an Xception Net classifier to detect Deepfakes with a greater degree of stability against subtle manipulation techniques often used in real-world environments. The generator can add perturbations that are controlled and perceptually invisible to the human eye so that the combination of both the models means that the generator will produce "hard" samples, and thus allow the deepfake detector to learn features about manipulating, while making more realistic predictions. Additionally, by completing the hybrid framework in two stages: baseline training and GAN augmented fine-tuning, the enhanced hybrid framework outperformed the baseline model and improved multiple metrics: Accuracy, F1, and ROC-AUC. Furthermore, the extensive evaluation of using external Deepfake samples and Grad-CAM Explainability provided confidence that this combination of models will significantly help in improving the performance of the next generation of Media Forensics Tools.

Future research can build on this work by including multimodal signals and temporal consistency models (e.g., Vision Transformers, 3D CNNs) to augment robust detection pipelines. In conjunction with stronger generative refiners (e.g., diffusion-based adversarial perturbation generators) and other technologies (e.g., Vision Transformers, 3D CNNS), this may increase the robustness to new and increasingly sophisticated methods of deepfake generation. By including social context cues and metadata-driven graph analysis as components of a detection pipeline, it will be possible to combine content forensics with propagation-based intelligence in an end-to-end detection framework.

8. References

[1] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," WIREs Data Mining and Knowledge Discovery, vol. 14, no. 2, p. e1520. [Online]. Available: https://doi.org/10.1002/widm.1520

[2] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, and L. Yuan, "DF40: Toward next-generation deepfake detection," in Proc. NeurIPS Datasets and Benchmarks Track, 2024. [Online]. Available: https://github.com/YZY-stack/DF40

[3] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: Challenges and opportunities," Artificial Intelligence Review, vol. 57, pp. 159–184. [Online]. Available: https://doi.org/10.1007/s10462-024-10810-6

[4] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, and V. Vimal, "Deepfake generation and detection: Case study and challenges," IEEE Access, vol. 11, pp. 143296–143315. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3342107

[5] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," Heliyon, vol. 11, p. e42273. [Online]. Available: https://doi.org/10.1016/j.heliyon.2025.e42273

[6] N. Bansal, T. Aljrees, D. P. Yadav, K. U. Singh, A. Kumar, G. K. Verma, and T. Singh, "Real-time advanced computational intelligence for deep fake video detection," Applied Sciences, vol. 13, no. 3,
p. 3095. [Online]. Available: https://doi.org/10.3390/app13053095

[7] Z. N. Ashani, I. S. C. Ilias, K. Y. Ng, M. R. K. Ariffin, A. D. Jarno, and N. Z. Zamri, "Comparative analysis of deepfake image detection method using VGG16, VGG19 and ResNet50," J. Adv. Res. Appl. Sci. Eng. Technol., vol. 47, no. 1, pp. 16–28. [Online]. Available: https://doi.org/10.37934/araset.47.1.1628

[8] A. Heidari, N. J. Navimipour, H. Dag, S. Talebi, and M. Unal, "A novel blockchain-based deepfake detection method using federated and deep learning models," Cognitive Computation, vol. 16, pp. 1073–1091. [Online]. Available: https://doi.org/10.1007/s12559-024-10255-7

[9] M. Astrid, E. Ghorbel, and D. Aouada, "Detecting audio-visual deepfakes with fine-grained inconsistencies," in Proc. British Machine Vision Conference (BMVC), 2024. [Online]. Available: https://bmva-archive.org.uk/bmvc/2024/papers/Paper_695/paper.pdf

[10] H. Khan, Z. Liu, Y. Li, and L. Song, "LGDF-Net: Local and global feature-based dual-branch fusion networks for deepfake detection," Applied Intelligence, 2024. [Online]. Available: https://doi.org/10.1007/s10489-024-05170-4

[11] S. Kumar, A. Gupta, R. Mishra, and T. Singh, "Deepfake detection techniques: A comparative survey," SSRN Preprint. [Online]. Available: https://doi.org/10.2139/ssrn.5240416

[12] V. Bansal, R. Kumawat, and S. Anand, "Faceswap Finder: A fusion-based deepfake detection technique," Procedia Computer Science, vol. 235, pp. 231–238. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.123

[13] K. Devi and S. Rajasekaran, "Deepfake video detection using AdaBoost on DFDC dataset," Procedia Computer Science, vol. 235, pp. 210–218. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.117

[14] H. Abbasi, F. Riaz, and A. A. Shah, "Comparative evaluation of deepfake detection using CNN architectures: Xception, ResNet, and VGG," Procedia Computer Science, vol. 235, pp. 198–209. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.115

[15] B. Chepchirchir and K. Mboli, "Comparative analysis of CNN, RNN, and Vision Transformers for deepfake detection," Procedia Computer Science, vol. 235, pp. 239–247. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.124

[16] P. Sharma and R. Rawat, "Enhancing deepfake detection through dynamics of facial expressions using masked autoencoders," Applied Sciences, vol. 15, no. 3, p. 1225. [Online]. Available: https://doi.org/10.3390/app15031225

[17] R. Tolosana, C. Rathgeb, A. Morales, and C. Busch, "Deepfake generation and detection: State of the art, open challenges, and future directions," Information Fusion, vol. 99, p. 101861. [Online].
Available: https://doi.org/10.1016/j.inffus.2023.101861

[18] Y. Zhang, J. Liu, and M. Chen, "Micro-expression dynamics for deepfake detection: A novel approach," Neural Computing and Applications, 2025. [Online]. Available: https://doi.org/10.1007/s00521-025-09876-9