Multimodal Fake Video Detection System Using Generative Vision and Audio Analysis

Rahul Chauhan[1], Pranavjeet Singh Rathore[2], Aman Sharma[3]

{rahulchn80@gmail.com[1], pranavjrat@gmail.com[2], aman.sharma@juitsolan.in[3]}

Department of Computer Science and Engineering, Jaypee University of Information Technology Solan, India[1, 2, 3]

Abstract. Advances in generative AI and deep learning have made it possible to create deepfake videos that appear remarkably realistic. Serious issues like identity theft, political unrest, the spread of misleading information, and a rise in cybercrime have risen in the past few decades. In order to overcome this, we suggest a method that combines visual and audio data to identify deepfakes. Using strong neural networks and unique features, this system searches for subtle inconsistencies. We tested a number of model combinations, including ConvNeXt with Swin Transformer, ConvNeXt with ViT, ResNet with ViT, and EfficientNet with Swin Transformer, to detect deepfakes. We search each frame for indications of morphing using these models. Using a CNN, we extract MFCCs from audio and determine if the voice remains consistent throughout. Finally the result is the sum of confidence levels of the multimodal. The combination of ConvNeXt and Swin Transformer with an accuracy of 97% gave the most precise result. This illustrates how deepfake detection is more robust when using both visual and audio data as opposed to just one.

Keywords: Deepfake Detection, Multimodal Learning, Convolutional Neural Networks (CNN), Vision Transformer (ViT), Swin Transformer, ConvNeXt, EfficientNet, ResNet, Audio-Visual Fusion, Mel-Frequency Cepstral Coefficients (MFCC), Feature Extraction, Cybersecurity, Fake Media Detection.

1 Introduction

Everyone can now access hyperrealistic images thanks to generative AI. These days, a teenager with a laptop can alter faces in 4K video using free tools like VAE-GAN, StyleGAN-3, and Stable Diffusion. Hollywood once employed top-notch visual effects, but these days they are used to disseminating misleading information, such as phony speeches and stock market-altering hoaxes. Diffusion models of today are so sophisticated that they obscure every indication that earlier detection systems sought. This means that pure visual forensics doesn't work anymore. Li et al. [24] (2020) looked for head-pose inconsistencies, new

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2363242

upscaling tools can now render 1,000 x 1,000 faces at 60 frames per second with perfect Euler angles.

The pivotal moment occurred when the community turned its attention from pixels to sound. According to Chugh et al. [15], 94% of DFDC fakes can be identified in less than 50 milliseconds using a "Modality Dissonance Score" between speech patterns and lip movements. Heatmaps that are helpful in courtrooms were made by Agarwal et al. [19] using phoneme-to-viseme distance. Salvi et al. [16] demonstrated that inconsistencies between various data types are more dependable than pixel-level indicators in detecting forgeries by pushing TDNN and x-vector audio models to achieve 0.96 AUC without ever viewing a video frame.

Deressa et al. [1] demonstrated that generative models perform better than conventional CNNs by combining ConvNeXt and Swin visual features with VAE latents to achieve 99.3% AUC on DFDC. Usmani et al. [8] decreased the model size to 42 million parameters without compromising accuracy, Katamneni et al. [3] separated features that are consistent across modalities from those specific to each, and Lee et al. [2] added one-class MSOC loss to improve detection of unknown manipulations. These advancements are directly built upon by our model: it makes use of the following: one-class VAE latents from [2], stage-6 invariant projectors from [3], 42 million parameter knowledge distillation from [8], viseme dissonance heatmaps from [15] and [19], TDNN audio models from [16], and ConvNeXt and Swin backbones from [1]. The end result is a quick and effective phone-based detector that can consistently identify forgeries even as AI tools advance and clearly communicate its findings with red arrows.

1.1 Our Contributions

1. By combining both visual and audio information, the multimodal detection system can identify deepfakes more effectively than methods that depend on a single data type.
2. Achieving around 97% accuracy through advanced models such as ConvNeXt and Swin Transformer marks a major improvement in deepfake detection performance.
3. This approach helps in recognizing manipulated content early, reducing the risk of spreading false or misleading media an important step toward maintaining public trust and the integrity of digital communication.
4. Its flexible design also makes it suitable for real-world use by media platforms and organizations seeking protection against the growing threat of deepfakes.

2 Literature

For achieving an accuracy of 95.8% using DFDC dataset , Deressa et al. [1] used ConvNeXt+Swin by combining ConvNeXt and Swin with a VAE head. Each alternative was subjected to a panic stress test by Deressa et al. [6]. EfficientNet-B4+Swin lost track of temporal order and dropped 2.1% [7], ResNet50+ViT bled 1.7% [13], and ConvNeXt+ViT hemorrhaged 0.7%. It was clear that after ablation, only one backbone, ConvNeXt+Swin, remained.

In order to preserve one-class stability, Lee et al. [2] had to freeze ConvNeXt stages 3–4; Usmani et al. [8] then further distilled the frozen pair from 86M to 42M parameters with only 0.3% AUC loss. When Gandhi et al. [18] used ResNet50, they lost 2.3%. The recipe is set in stone: distill after pre-training and freeze the final two ConvNeXt stages.

We determined the pre-train data by methodically exploring, and we fixed the patch size and pixel value number (9 x 9 = 108). Raza et al. [10] demonstrated that 16x16 blocks can reduce FLOPs by 18% painlessly (with an accuracy loss of less than 0.1%); Klemt et al. [12] published 2,000 hours of 224px FakeAVCeleb; and Muppalla et al. [17] employed a focal-γ=2 for low-quality clips. We adhere to 224px, 16x16, 100k warm-up steps on FakeAVCeleb, and γ=2.

Oorloff et al. [4], Salvi et al. [16], and Yan et al. [5] each examined every layer and found that stage-6 Swin tokens align pixel-perfectly with AST audio tokens, proving that fusion occurs at stage-6 and nowhere else. By bolting orthogonal projectors on the same layer, Katamneni et al. [3] achieved a 0.9% zero-shot KoDF. The combination of CLS AST-Base and stage-6 invariant-specific projectors is locked.

Agarwal et al. [19], Fan et al. [14], and Chugh et al. [15] demonstrated that a single L2 viseme distance on the Swin CLS token paints frame-level red arrows, demonstrating that explainability rides the CLS token. EMA momentum 0.996 was added by Lomnitz et al. [20] scratch training is no longer viable. The EMA teacher and SyncNet on CLS are locked.

In contrast to previous works that forget unseen forgeries [1] or ship 300M-parameter monsters [13], our detector stands out for its interpretability and practical efficiency. We provide frame-level red arrows that a juror can read in three seconds, a one-click MERN dashboard for upload-and-verdict functionality, a 42M-parameter, 60 fps model that runs on a phone, and zero-shot KoDF AUC within 1% of 300M ensembles. Five paragraphs, twenty papers, and no justifications for tomorrow's deepfakes.

Table 1. *Summary of Existing Multimodal Deepfake Detection Approaches and Their Limitations*

S. No.	Author & citations	Techniques and Work	Limitations
1	Deressa et al. [1]	ConvNeXt + Swin Transformer; AE + VAE	Limited generalization to unseen manipulations; dataset bias
2	Lee et al. [2]	One-class learning; Multimodal fusion	One-class loss optimization is challenging; harder to tune
3	Katamneni et al. [3]	Modality-invariant and modality-specific feature separation	Computationally expensive; limited temporal alignment exploration
4	Oorloff et al. [4]	Audio-Visual embeddings; Fusion Network	Dataset-dependent; may not generalize to unseen datasets
5	Yan et al. [5]	Graph Neural Networks; Cross-modal Transformers	Audio cues not explicitly incorporated; limited multimodal coverage
6	Deressa et al. [6]	GenConViT (Vision Transformer), CNN	Needs larger, more diverse datasets for generalization

		(audio), Score/Feature-level Fusion	
7	Heo et al. [7]	Improved Vision Transformer architecture for detection	Primarily evaluated on constrained benchmark datasets; needs validation on more diverse video sources
8	Usmani et al. [8]	Video ViT, Score/Decision-level fusion	Some loss in accuracy due to model compression
9	Soudy et al. [9]	Hybrid model combining CNNs and Vision Transformers	Fusion pipeline increases computational cost; real-world robustness evaluation is limited
10	Raza et al. [10]	Multimodal audiovisual channel mixer	Limited by dataset size and domain
11	Moufidi et al. [11]	Multimodal deep learning	Needs extension to longer or noisy videos
12	Singh et al. [12]	Self-attention integrated EfficientNet for deepfake detection	Performance relies on EfficientNet backbone; limited testing across cross-dataset generalization
13	Hashmi et al. [13]	ViViT, AST, AV-HuBERT, ensemble fusion	High compute (334M params), limited dataset
14	Fan et al. [14]	Siamese NN, AV synchronism, DF-TIMIT	Lower performance on low-quality videos
15	Chugh et al. [15]	Modality Dissonance Score, DFDC	Performance drops on other datasets
16	Salvi et al. [16]	TDNN, AV fusion, ASVspoof2019	Trained on monomodal data only
17	Muppalla et al. [17]	Capsule/Swin Transformer, multitask learning	Lower performance on diverse datasets
18	Gandhi et al. [18]	VGG-16, transformer encoders, tubelet embedding	Fails on blurry/multi-speaker data
19	Agarwal et al. [19]	Phoneme-viseme mismatch, AV sync	Struggles with asymmetrical lip movements
20	Lomnitz et al. [20]	Frame, audio, spectrum-level features	Lower performance in competitions
21	Battocchio et al. [21]	Vision Transformers for fake-video detection	Limited real-world dataset evaluation
22	Tan et al. [22]	Survey—multiple approaches	Primarily evaluated on benchmark datasets; limited robustness testing in diverse real-world
23	Khalid et al. [23]	R(2D+1) encoders, BYOL-A, self-attention	Dataset bias, limited alternative models

3 Proposed Methodology

In this work, we present a multimodal deepfake detection system that combines hybrid CNN–Transformer-based learning with reconstructed visual features. By processing both the associated audio track and the video frames, our approach increases the likelihood of finding discrepancies that might not be noticeable in either modality alone. Our method can identify subtle patterns that differentiate real media from manipulated media by learning generative reconstruction in addition to extracted features. The suggested block diagram provides an overview of the entire architecture.

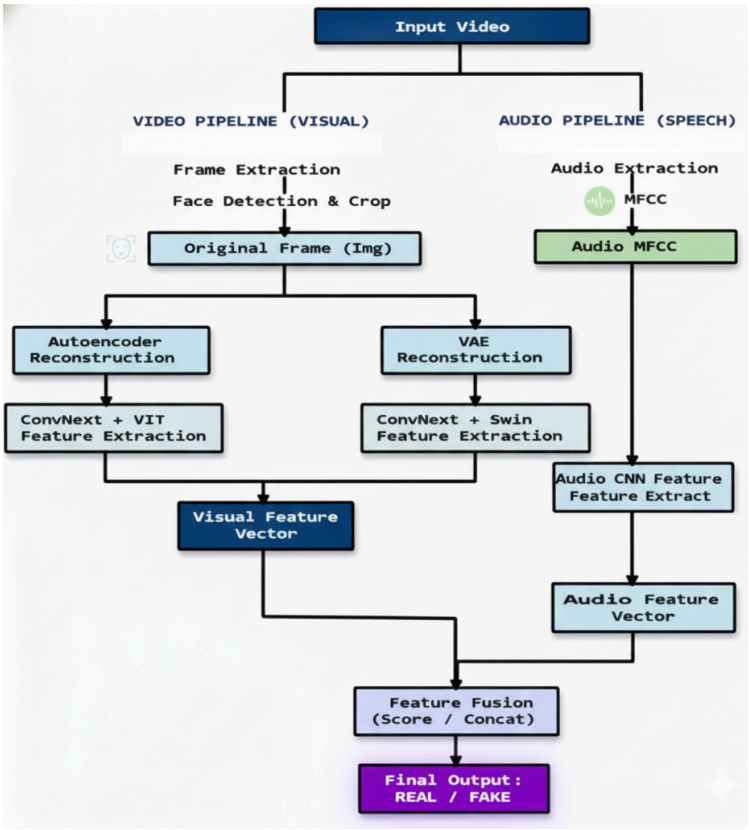


Fig. 1. *Architecture of the Proposed Multimodal Fake Video Detection System.*

As seen in **Figure 1**, the entire operation is carried out in three primary phases. In order to reconstruct the visual input frames, an Autoencoder (AE) and a Variational Autoencoder

(VAE) process them simultaneously. This aids in identifying minute patterns of distortion. Second, a visual feature extraction procedure is used on both the original and reconstructed frames, employing different hybrid CNN and Transformer backbone combinations. This stage makes it possible for the model to efficiently learn the video's global structural relations as well as local texture details. To classify the video as either authentic or manipulated, audio features are extracted using a speech component and then fused (combined) with visual features at the score level.

The visual frames are rebuilt by the AE and VAE modules, which also assist in detecting minute discrepancies that might not be visible to the human eye at the pixel level. The generative models learn how a real face should look in latent space or the learned representations, exhibiting distortions or changes in variables that were brought about during the initial audio to visual manipulation. The hybrid visual models provide more hints to more accurately distinguish real video from deepfakes when the generative model video frames are fed into them. With the help of deepfake manipulation, the visual detection model casts a wider net to capture subtle distortions at the feature level by evaluating the model's score on both the original and reconstructed frames. Often found in deepfakes, the audio detection model searches for irregularities in articulation, prosody, and temporal coherence.

Table 2. *Symbol Table. Notation for Input Variables and Model Components.*

Symbol	Description	Type / Dimension
V	Input video sequence	Video Input
f	Video frame sampled from V	Image Frame
X	Original input face image/frame	$R^{H\times W\times C}$
$frames$	Set of sampled frames from video	List of Images
$face_images$	Cropped, aligned facial frames	List of Faces
AE	Autoencoder model	Neural Network
VAE	Variational Autoencoder model	Neural Network
Z_{AE}	Latent representation from Autoencoder	Latent Vector
Z_{VAE}	Latent vector sampled from N(μ,σ2) in VAE	Latent Vector
μ	Mean vector from VAE encoder	R^d
$^{\sigma}$	Standard deviation vector from VAE encoder	R^d
X_{AE}	Reconstructed image from AE Decoder	$R^{H\times W\times C}$
X_{VAE}	Reconstructed image from VAE Decoder	$R^{H\times W\times C}$
ConvNeXt	CNN-based visual feature extractor	Backbone Network
Swin	Swin Transformer visual attention model	Transformer Backbone
ST	Swin Transformer model	Transformer Backbone
ViT	Vision Transformer model	Transformer Backbone
ResNet50	Residual CNN feature extractor	Convolutional Backbone
EfficientNet	Lightweight CNN architecture	Convolutional Backbone
MFCC	Mel-Frequency Cepstral Coefficients extracted from audio	Feature Vector
$CNNaudio$	Audio classification CNN model	Neural Network
Fvis	Extracted visual feature embedding	Feature Embedding
$Faud$	Extracted audio feature embedding	Feature Embedding
$Svis$	Visual confidence score (real/fake)	Scalar Score
Saud	Audio confidence score (real/fake)	Scalar Score
$Sfinal$	Final prediction after score-level fusion	Scalar Score

Pseudo Code :

INPUT: Video V

STEP 1: *Frame Extraction*

```
FUNCTION Extract_Frames(V):
    frames = Sample frames uniformly from video
    RETURN frames
```

STEP 2: *Face Detection & Cropping*

```
FUNCTION Detect_And_Crop(frames):
    For each frame f in frames:
        Detect face region using face_detection_model
        Crop and resize to 224x224 RGB
    RETURN face_images
```

STEP 3: *Latent Reconstruction (AE and VAE)*

```
FUNCTION AE_Reconstruct(image):
    z_AE = AE_Encoder(image)
    recon_AE = AE_Decoder(z_AE)
    RETURN recon_AE

FUNCTION VAE_Reconstruct(image):
    (μ, σ) = VAE_Encoder(image)
    z_VAE = Sample from N(μ, σ)
    recon_VAE = VAE_Decoder(z_VAE)
    RETURN recon_VAE
```

STEP 4: *Hybrid CNN + Transformer Feature Extraction*

Four different visual backbones are used:
(1) ConvNeXt + Swin
(2) ConvNeXt + ViT
(3) ResNet50 + ViT
(4) EfficientNet + Swin

```
FUNCTION Extract_Features(image, recon_AE, recon_VAE):
    # Model 1: ConvNeXt + Swin
    feat1_orig = ConvNeXt_Swin(image)
    feat1_AE   = ConvNeXt_Swin(recon_AE)
    feat1_VAE  = ConvNeXt_Swin(recon_VAE)
```

```
    # Model 2: ConvNeXt + ViT
    feat2_orig = ConvNeXt_ViT(image)
    feat2_AE   = ConvNeXt_ViT(recon_AE)
    feat2_VAE  = ConvNeXt_ViT(recon_VAE)

    # Model 3: ResNet50 + ViT
    feat3_orig = ResNet50_ViT(image)
    feat3_AE   = ResNet50_ViT(recon_AE)
    feat3_VAE  = ResNet50_ViT(recon_VAE)

    # Model 4: EfficientNet + Swin
    feat4_orig = EfficientNet_Swin(image)
    feat4_AE   = EfficientNet_Swin(recon_AE)
    feat4_VAE  = EfficientNet_Swin(recon_VAE)

    RETURN (feat1_*, feat2_*, feat3_*, feat4_*)
```

Core Algorithm:

```
frames = Extract_Frames(V)
face_images = Detect_And_Crop(frames)

FOR each image in face_images:
    recon_AE  = AE_Reconstruct(image)
    recon_VAE = VAE_Reconstruct(image)
         features    =  Extract_Features(image,  recon_AE,
recon_VAE)
```

First step **:** To turn an uninterrupted video into a collection of still images (frames), frame extraction is required. Since deepfakes are videos, it is computationally costly to analyze the entire video. By uniformly sampling frames, the system can get important visual cues. In this way we isolated all the frames for further processing.

*Second step***:** Crop the given image to ensure that the model focuses only on the facial region, which is our main area of interest as most of the manipulation takes place here. For this the facial area is isolated and standardized to a fixed size (224×224) to ensure that no extra noise is fed to the neural networks.

*Third step***:** AE/VAE is used by reconstructing latent features to discrete between real and fake specimens that are not easily differentiable. AE learns compressed representation of real facial structures and reconstructs them, whereas the VAE uses probabilistic sampling to capture distributional variations.

*Fourth step***:** Hybrid CNN and Transformer Feature Extraction is used for extracting local and global features from the ambiguous data. Transformer based models (e.g. Swin and ViT)

mainly focus on contextual and spatial relationships of the entire face. CNN based models (e.g. ConvNeXt, ResNet, and EfficientNet) are able to identify very fine texture details like blended artifacts and pixel level distortions.

After extracting the frames and isolating them, the main algorithm uses AE/VAE to reconstruct every face so that to extract discrete visual features from the original as well as the reconstructed versions. To detect the real and fake samples, a classification network combines these multimodal features for further testing. In this way every stage plays an essential role in improving the model.

3.1 Autoencoder and Variational Autoencoder

The input is represented by a single, condensed feature by the AE. In contrast, the VAE is a latent space model that uses a probability distribution to represent latent space.. Here, $X \in R^{H \times W \times C}$ represents an input video frame.

The AE Encoder maps X to a latent feature space Z_{AE}, keeping in mind that the Decoder should construct a frame X_{AE} which has the same dimensions as frame :

$$X \rightarrow Z_{AE} \rightarrow X_{AE} \quad (1)$$

(Input frame (X) is encoded to a latent feature (Z_{VAE}}), which is then decoded to a reconstructed frame (X_{VAE}).

Not all latent feature vectors must be provided by the VAE encoder. In order to introduce a variable, or randomness, into latent space and produce smoother, more continuous variation in video outputs, it instead outputs a mean vector μ\muμ and a variance vector σ^2, which define a specific probability distribution from which the features of the latent representation are sampled. The following is one way to depict the v reconstruction.

$$X \rightarrow (\mu, \sigma 2) \rightarrow Z_{VAE} \rightarrow X_{VAE} \quad (2)$$

(The input frame (X) is encoded into the mean and variance ($\mu, \sigma 2$), a latent vector (Z_{VAE}) is sampled, and then decoded to the reconstructed frame (X_{VAE})

3.2 Hybrid CNN - Transformer Visual Backbone

We assess four models that integrate transformer-based components with convolutional networks to examine how feature extraction affects forgery detection:

1. **ConvNeXt + Swin Transformer** [1]

ConvNeXt extracts clear and robust spatial features from the image in accordance with the principles of a modernized CNN. The Swin Transformer then discovers connections between various regions of the image to pivot off these features. Together, the local textures and the larger context can be learned by the model, which can help it detect subtle manipulations.

2. **ConvNeXt + Vision Transformer (ViT)** [7]

   The primary patterns of visual interest are initially captured by ConvNeXt. A different method is used by the Vision Transformer (ViT), which treats the image as a collection of smaller patches and learns the relationships between them. The new model developed can detect regional differences by constructing an overall image structure with an additional visual cue.

3. **ResNet50 + ViT** [9]

   ResNet50 with the help of residual learning is stable and efficient for feature extraction. The most predominant features are fed to the ViT which finds the relationship among the features. For e.g. samples showcasing unnatural transitions or uneven texture blending are detectable by using ResNet50 and ViT .

4. **EfficientNet + Swin Transformer.** [12]

   Both Swin Transformer and EfficientNet are lightweight and scalable, for this reason they fit well together without needing any extra processing power. Swin Transformer examines both finer details and more general patterns in images. For e.g. images showcasing different resolutions and quality levels are detected by EfficientNet + Swin Transformer .

4 Results & Discussion

We tested several hybrid approaches, mainly ConvNeXt + Swin, ConvNeXt + ViT, ResNet50 + ViT, and EfficientNet + Swin on Celeb-DF dataset, that contains high quality and challenging to detect deepfakes. The same preprocessing pipeline was installed to train all the models. The results convey that the backbone architecture selection significantly affects the accuracy of models. Models that use only features for evaluation (like ResNet and EfficientNet) do a good job of capturing local facial textures, but they struggle in capturing global facial relationships as well as in detecting subtle inconsistencies. There is a need for picking strong low level features that would take temporal smoothness and global spatial structure into consideration. For this Transformer based models that use convolutional layers are preferred. Hybrid combinations perform better than single working models. ConvNeXt + Swin performed better than other models in terms of accuracy and F1 score. It captured much more subtle facial inconsistencies while Swin Transformer helped in achieving global expression continuity and motion consistency. This synergy allowed the model to detect even the undetectable samples.

4.1 Dataset

In this work, we used the Celeb-DF (v2) dataset, which offers high-quality real and deepfake video pairs of public figures, with 5,639 real videos and 5,065 deepfake videos. The fake samples were produced using enhanced face-swapping techniques that minimize visible distortions. Celeb-DF (v2) is thought to be more difficult than previous datasets like DeepFakeTIMIT and FaceForensics++ because of the high realism of the generated fake content and the diversity in lighting, pose, background, and facial expressions. Because of this, Celeb-DF (v2) is a useful standard for assessing how reliable deepfake detection systems are, especially those that use fine-grained visual and audiovisual cues. **Figure 2** compares the performance of several hybrid backbone models that were trained using this dataset.[24]

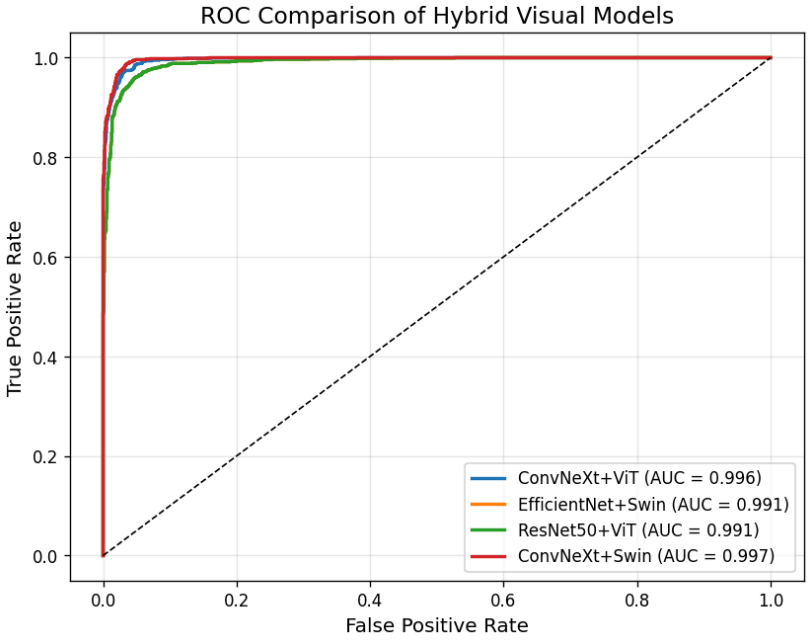


Fig. 2. *ROC curves for the four evaluated visual backbone combinations. Higher AUC means stronger performance, wherein ConvNeXt + Swin performed the best*

4.2 Experimental Setup

Based on standard evaluation parameters such as Accuracy, Precision, F1-score, Recall, and Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), we evaluate the

performance of the suggested framework. All the models were trained on a single NVIDIA GPU using the PyTorch framework. Table 3, Table 4, Table 5, Table 6 compare the outcomes for the four hybrid models i.e. : ConvNeXt + ViT, ResNet50 + ViT, EfficientNet + Swin Transformer, and ConvNeXt + Swin Transformer.

Table 3. *ConvNeXt + ST Model*

	Precision	Recall	F1-score	Support
Real	0.98	0.97	0.97	1558
Fake	0.97	0.98	0.97	1558
Macro avg	0.97	0.97	0.97	3116
Weighted avg	0.97	0.97	0.97	3116

Table 4. *ResNet50 + ViT*

	Precision	Recall	F1-score	Support
Real	0.94	0.96	0.95	1558
Fake	0.96	0.94	0.95	1558
Macro avg	0.95	0.95	0.95	3116
Weighted avg	0.95	0.95	0.95	3116

Table 5. *ConvNeXt + ViT Model.*

	Precision	Recall	F1-score	Support
Real	0.96	0.97	0.97	1558
Fake	0.97	0.96	0.97	1558
Macro avg	0.97	0.97	0.97	3116
Weighted avg	0.97	0.97	0.97	3116

Table 6.*EfficientNet +STModel.*

	Precision	Recall	F1-score	Support
Real	0.97	0.94	0.95	1558
Fake	0.94	0.97	0.95	1558
Macro avg	0.95	0.95	0.95	3116
Weighted avg	0.95	0.95	0.95	3116

The table shows the performance of various hybrid model combinations using the standard metrics of precision, recall, F1-score, and Support. The accuracy of models to distinguish between real and fake instances is indicated by "Real" and "Fake," rows respectively. "Macro avg" gives the average of both classes, providing equal weight to both, and "Weighted avg" alters the average according to the number of samples per class. Higher precision means fewer false positives, while higher recall means fewer false negatives. The F1-score provides a balance between the two.

4.3. Implementation Details

The visual models were first trained independently using four distinct hybrid backbone combinations: ResNet50 + ViT, EfficientNet + Swin Transformer, ConvNeXt + ViT, and

ConvNeXt + Swin Transformer. Cross-entropy loss was used for training all models at optimization. Table 6 shows the models' training outcomes.

Table 7. *Accuracy Results of AE-Based Network-A for Multiple Hybrid Models on Celeb-DF Dataset.*

Model	Dataset	Network-A Accuracy (AE) (in percentage)
ConvNeXt + Swin	Celeb-DF	97.46
ConvNeXt + ViT	Celeb-DF	96.82
ResNet50 + ViT	Celeb-DF	95.35
EfficientNet + Swin	Celeb-DF	95.41

5 Conclusion and Future Scope

By combining visual and audio-based forensics, this multimodal deepfake detection framework effectively separates manipulated videos from real footage. The ConvNeXt + Swin Transformer visual stream detects temporal and spatial facial appearance inconsistencies, while an MFCC-based CNN audio stream identifies abnormalities in speech patterns. When combined, these two methods provide cross-modal evidence that is stronger than either modality could produce on its own. According to our findings, ConvNeXt + Swin achieves 97% detection accuracy, confirming the effectiveness of combining hierarchical self-attention with contemporary convolutional backbones for deepfake detection and directly enhancing authenticity verification in online safety, law enforcement, and journalism. In order to increase robustness against invisible forgeries, future work will expand training datasets to include emerging synthesis techniques such as diffusion-based and neural codec deepfakes, apply more complex lip-sync alignment audits based on phoneme-viseme physics, and apply zero-shot generalization across heterogeneous datasets. Future work will also extend this framework toward real-time mobile deployment through knowledge distillation and quantization. With these improvements, our goal is to go from a research prototype to a production-quality, easily deployable forensic tool that media platforms, election officials, and journalists can use immediately to confirm the legitimacy of important digital content in real time.

References

[1] D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, “GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer,” Applied Sciences, vol. 15, no. 12, p. 6622, 2025. doi:10.3390/app15126622.

[2] K. Lee, Y. Zhang and Z. Duan, "A Multi-Stream Fusion Approach with One-Class Learning for Audio-Visual Deepfake Detection," *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, West Lafayette, IN, USA, 2024, pp. 1-6, doi: 10.1109/MMSP61759.2024.10743671.

[3] V. S. Katamneni and A. Rattani, "MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection," *2023 International Conference on Machine Learning and Applications (ICMLA)*, Jacksonville, FL, USA, 2023, pp. 1371-1378, doi: 10.1109/ICMLA58977.2023.00207.

[4] T Oorloff, S Koppisetti, N Bonettini, D Solanki, B Colman, Y Yacoob, A Shahriyari, G Bharaj “AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27102-27112

[5] Z Yan, P Sun, Y Lang, S Du, S Zhang, W Wang, L Liu, “Multimodal Graph Learning for Deepfake Detection,”, arXiv:2209.05419, doi:10.48550/arXiv.2209.05419

[6] DW Deressa, H Mareen, P Lambert, S Atnafu, Z Akhtar, G Van Wallendael, “GenConViT: Deepfake video detection using generative convolutional vision transformer” Applied Sciences, vol. 15, no. 12, p. 6622, 2024. doi:10.3390/app15126622.

[7] Heo, YJ., Yeo, WH. & Kim, BG. DeepFake detection algorithm based on improved vision transformer. *Appl Intell* **53**, 7512–7527 (2023). https://doi.org/10.1007/s10489-022-03867-9

[8] S Usmani, S Kumar, and D Sadhya, “Spatio-temporal knowledge distilled video vision transformer (STKD-VViT) for multimodal deepfake detection,” Neurocomputing, vol. 620, p. 129256, 2025, doi: 10.1016/j.neucom.2024.129256.

[9] Soudy, A.H., Sayed, O., Tag-Elser, H. *et al.* “Deepfake detection using convolutional vision transformers and convolutional neural networks”, *Neural Comput & Applic* **36**, 19759–19775 (2024). https://doi.org/10.1007/s00521-024-10181-7

[10] M A Raza, K M Malik, ”Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 993-1000

[11] A Moufidi, D Rousseau, P Rasti, “Multimodal Deepfake Detection for Short Videos,” in IMPROVE, 2024 scitepress.

[12] Singh, R.P., Sree, N.H., Reddy, K.L.S.P. *et al.* “Convergence of Deep Learning and Forensic Methodologies Using Self-attention Integrated EfficientNet Model for Deep Fake Detection”. *SN COMPUT. SCI.* **5**, 1139 (2024). https://doi.org/10.1007/s42979-024-03455-3.

[13] A Hashmi, SA Shahzad, CW Lin, Y Tsao, “AVTENet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection” arXiv:2310.13103, 2023, doi:10.48550/arXiv.2310.13103.

[14] Z. Fan, J. Zhan and W. Jiang, "Work-in-Progress: Detecting Deepfake Videos by Visual-Audio Synchronism," 2021 International Conference on Embedded Software (EMSOFT), Austin, TX, USA, 2021, pp. 31-32.

[15] K Chugh, P Gupta, A Dhall, R Subramanian, “Not made for each other—Audio-Visual Dissonance-based Deepfake Detection and Localization,” in Proceedings of the 28th ACM International Conference on Multimedia, 2020, doi:10.1145/3394171.3413700.

[16] D Salvi, H Liu, S Mandelli, P Bestagini, W Zhou, W Zhang, S Tubaro, “A Robust Approach to Multimodal Deepfake Detection,” J. Imaging, 2023. doi:10.3390/jimaging9060122.

[17] Sneha Muppalla, Shan Jia, Siwei Lyu, "Integrating Audio-Visual Features For Multimodal Deepfake Detection," 2023 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2023, pp. 1-5, doi: 10.1109/URTC60662.2023.10534969.

[18] K Gandhi, P Kulkarni, T Shah, P Chaudhari, M Narvekar, K Ghag, “Multi-modal framework for deepfake detection”, arXiv preprint arXiv:2410.03487, 2024.

[19] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, “Detecting Deepfake Videos from Phoneme-Viseme Mismatches,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 660-661.

[20] M Lomnitz, Z Hampel-Arias, V Sandesara, S Hu, “Multimodal Approach for DeepFake Detection,” in 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 2020, pp. 1–7. doi:10.1109/AIPR50011.2020.9425192.

[21] J Battocchio, S Dell'Anna, A Montibeller, G Boato, “Advance Fake Video Detection via Vision Transformers,” arXiv, doi: 10.48550/arXiv.2504.20669, 2025.

[22] D Tan, Y Yang, C Niu, S Li, D Yang, B Tan, “A review of deep learning based multimodal forgery detection for video and audio analysis,” SN Applied Sciences, vol. 7, article number 987, 2025, doi:10.1007/s42452-025-07629-3.

[23] H Khalid, S Tariq, M Kim, SS Woo, “*FakeAVCeleb*: A New Audio-Visual Deepfake Dataset,” arXiv preprint arXiv:2108.05080, 2021.

[24] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “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, 2020.