Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 20-07-2026; Accepted 09-11-2026; Published online 16-11-2027

Enhancement of Deepfake Identification Using Spatial, Temporal, Frequency and Audio-Visual Fusion Techniques

Jayapradha J jayapraj@srmist.edu.in
Department of Computing Technologies,
School of Computing Faculty of Engineering and Technology
SRM Institute of Science and Technology, Kattankulathur–603203 Tamil Nadu, India

Su-Cheng Haw sucheng@mmu.edu.my
Faculty of Computing and Informatics Multimedia University Jalan Multimedia
63100, Cyberjaya, Malaysia

Senthil Kumar T senthilt2@srmist.edu.in
Department of Computing Technologies,
School of Computing Faculty of Engineering and Technology
SRM Institute of Science and Technology, Kattankulathur–603203 Tamil Nadu, India

Malathi S jj_mals@eee.sastra.edu
Department of Electrical and Electronics Engineering,
SASTRA University, Thanjavur, India

Nilanjana Bhattacharya nb2168@srmist.edu.in
Software Development Engineer-Level 1,
JP Morgan Chase and Co, Hyderabed, India

Corresponding Author: Su-Cheng Haw

Copyright © 2026 Jayapradha J, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Rapid advancements in deepfake generation technology have met challenges in media. Dependability truthfulness authentication. Therefore, this paper presents a multi-modal deepfake detection anti-cheat system: spatial, temporal, frequency, and audiovisual synchrony features were incorporated. Visualization of ResNet18 pre-trained on imagenet data pounds down pixel-level features in the spatial channel, while a Long Short-Term Memory (LSTM) recurrent neural network in the temporal channel captures sequential behavior. Frequency artifacts are extracted using Fast Fourier Transform (FFT) and audiovisual synchrony compared by Dynamic Time Warping (DTW) between Mel Frequency Cepstral Coefficients (MFCC, audio) and lip movement locations (visual). In the last step, we unify spatial, temporal, frequency, and audio–visual synchrony features. The four feature streams are concatenated into a 1,153-dimensional vector for classification. The manuscript reports accuracies of 78.4%, 82.1%, and 79.6% for the spatial, temporal, and frequency models, and 91.3% for the fused model, with precision 0.901, recall 0.918, and F1-score 0.910 for fusion.Experimentally, a model is proposed for real-world testing that applies YouTube data as input to compare real and manipulated media without time limits. With multi-modal deep-

Citation: Jayapradha J, et al. Enhancement of Deepfake Identification Using Spatial, Temporal, Frequency and Audio-Visual Fusion Techniques. **Advances in Artificial Intelligence and Machine Learning**. 2026. [Ahead of Print]. https://dx.doi.org/10.54364/AAIML.2026.66358

fake detection framework, cross spatial, temporal, frequency and audio-visual synchrony detection is of a higher-class stability and reduces the misjudgments rate.

Keywords: ResNet18, Long short-term memory, Dynamic time warping, Mel frequency cepstral coefficients, Spatial, Temporal, Frequency, Deepfake.

1. INTRODUCTION

New opportunities and challenges are emerging because of deep-fake technology [1]. It has brought more interesting applications to entertainment and academic fields, and it has raised safety and ethical issues. Deepfakes use innovative machine learning algorithms to create fake videos and audio that are created overnight [2, 3]. These fabrications must be detected before damage occurs. This is true not only for news dissemination but also for court hearings and an individual's right to privacy. In the AI rebellion, a computer science man teamed up with a clever enemy and sprang up this immoral act of expertise: Generative Adversarial Networks (GANs) [4]. GANs can create fake stories using two neural networks, one of which creates one story and the other determines whether it is true or false [5, 6]. Thus, the effects of this recurs formation process were to continually produce something that was increasingly realistic and difficult for humans to decide whether it is true or false media. Deepfake technology was initially introduced in connection with its application to animatronics, video gaming proposal and Virtual Reality (VR) [7], but its potential harm had already been on center stage. Of malfeasance users began to manipulate it not only to make up fake news but also to prejudice people, and still further to unscrupulously cyber-punk the economic markets. This highlights the necessity for better detection methods. So now its accusation on deepfake technology is amplification of widely held social beliefs, the loss of your integrity and then who will believe in what in digital media. We are in a time that visual and audio evidence can be deemed irrefutable, and to create such content shatters faith, if not trust into truth and accountability. The seriousness of this issue has been acknowledged by administrate, organizations and scientists globally who have invested significant efforts for developing measures against it [8]. Deepfakes are being produced at break-neck speed with detectors having to stay one step ahead of the counterfeiting innovation, which is a very tough ask. This statement provides an end-to-end detection pipeline and makes delegates of the active fight against deep-fake spread due to the war, A modified detection model is also proposed and evaluated on the dataset that contains both original and fake videos collected in our research. It strives to supplement and revise current detection systems by focusing on preprocessing techniques and creating a suited-sized CNN that fulfills industry demands [9]. This latter half of the agenda will ensure that the proposed system, while pursuing high accuracy, is also practical and thereby functional for practical use. Since this research is meant to defend against deepfakes and keep the virtual world from falling into disrepute. In FIGURE 1, we present a high level of enabling detail on how the proposed model was constructed.

2. RELATED WORKS

The impressive rise of deepfake technology has sparked considerable interest in developing effective methods of detection. Over the years, researchers have found a myriad range of solutions to detect interfered media, and each one of them has solved definite gaps; however, they have

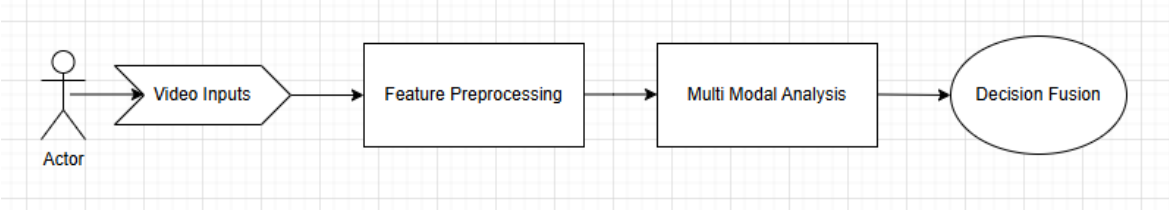


Figure 1: Layers in the Detection System

definite weaknesses that require enhancement. It has also been found that [10] temporal coherence analysis can be an effective technique in the recognition of deepfakes by examining the differences in frame sequences across time. The observation of motion dynamics and temporal artifacts has been performed with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models. These techniques are largely effective at detecting unusual discontinuities between video frames, mostly in synthesized material that does not have real temporal continuity. Despite being very effective, the temporal coherence strategies tend to overlook other modalities, including audio-visual discrepancies or high-frequency artifacts, that are also important to be detected accurately.

Specific interest has also been placed on audio-visual analysis [11], to identify misalignment between lip movement and the audio track of deepfaked videos. Mel-spectrograms [12] and lip synchronization models are quite useful in recording the differences between audio and visual streams. Such methods are most appropriate in identifying deepfakes that are produced with an incompatible audio file. The techniques are limited in their scope, however, because they lack spatial, temporal, and frequency-domain analysis and, therefore, miss important cues that would improve the detection rates.

The frequency domain has also been discussed in [13] as a good resource to detect forensic deepfakes. Synthesis artifacts (high-frequency artifacts) can disclose manipulation which would otherwise stay undetected either spatially or temporally. Such artifacts have been analyzed by methods like Fast Fourier Transform (FFT) [14], which detects anomalies in the frequency spectrum of manipulated video. Frequency-domain methods prove useful when applied but are inadequate alone because they ignore the elements of temporal coherence, behavioral inconsistency, and audio-visual inconsistencies, which are critical to full detection.

The shortcomings of single methods can be overcome by recent developments in multi-methodology integration. For example, a few studies have investigated the application of Capsule Networks [15] together with adversarial training to improve the robustness of detection models to change deepfake technologies. As spatial hierarchy-preserving networks, Capsule Networks are demonstrated to be highly efficient in detecting face manipulations that are rather subtle. When trained adversarial, such models can learn to generalize new deepfake methods. The nature of such advanced methodologies, however, ignores the value of multi-modal data integration, e.g., temporal coherence, audio-visual synchronization, and behavioral biometrics that are essential to state-of-the-art detection.

One of the first successful ways of detecting deepfakes was trying to extract spatial features using the technique of Convolutional Neural Networks (CNNs). In [16], the authors analyzed the performance of CNNs when detecting anomalies and pixel inconsistency in facial texture patterns. CNNs are

adept at learning hierarchical spatial representations which encode fine details that differ between real and manipulated content. However, CNN-based approaches are in the end constrained by an inability to learn temporal inconsistency or to generalize to unknown deepfake manifestations. To overcome the drawbacks of spatial-only approaches, recurrent neural networks (RNNs) have been used for temporal analysis, especially Long Short-Term Memory (LSTM) networks. In [17], an RNN was explored for identifying deepfake-specific temporal inconsistencies by modelling sequential dependencies in video frames. The model extracts feature from each frame using CNNs and feeds each frame's features into an LSTM, where the model learns motion patterns over time. This model can detect temporal inconsistencies that are not generally achievable using purely spatial analysis techniques.

The research presented a detection method [18] that proposed a new detection paradigm based on frequency-domain analysis and principles of factchecking. Whereas current methods of supervised classification rely on learning, the FACTOR framework analyzes deepfake inconsistencies through known attributes versus observed features. The authors exploit the basic limitation of deepfake generators, which attempt to create high-fidelity syntactic components, by analyzing contradictions of identity, speech, and appearance [19]. By evaluating identity, speech, and appearance attributes against known ground truth to detect zero-day deepfakes and address a rapidly evolving trend toward generative, synthetic deepfakes, the FACTOR approach provides adaptability and versatility.

The hybrid model investigated in [20, 21] combined ViTs with EfficientNet CNNs [22] as a detector. So Why are we not using ViTs EfficientNet is an efficient feature extractor, preserving local texture information and detail. Unlike ViTs that model long-range dependencies across the image using self-attention mechanisms. It achieved AUC: 0.951 and F1 Score: 88.0% outperforming CNN-based methods on DFDC (DeepFake Detection Challenge) dataset. For the case of combining local convolutional feature extraction with transformer-based representation learning, the robustness of the model to adversarial manipulations was enhanced. In [23], this paper dealt with localized deepfake manipulations which mean certain parts of a video or audio are altered. Localised Audio-Visual DeepFake [24] (LAVDF) :The authors proposed this dataset to assist in assessing the effectiveness of diverse deepfake detection frameworks by means of constraining forgery positioning temporally. In Ref. [25], the researchers trained a 3D-CNN architecture called Boundary Aware Temporal Forgery Detection (BA-TFD) and investigated the usage of contrastive learning and frame classification loss functions to recognize forged areas. In turn, this may allow for small content-transforming changes to the proposed model that binary classifiers cannot recognize or see.

The FaceForensics dataset, which [26] conducted in the article, is one of the first steps to deepfake standards. The paper analyzed various deepfake techniques including DeepFakes, Face2Face, FaceSwap and NeuralTextures. Some of the candidates trained with different compression levels, such as by speeding-up or slowing-down these modes during training had low test accuracies hence implying that domain information is adequate for accurate detection. The database has been periodically updated and is still a key benchmark for testing contemporary deepfake detection techniques. At the same time, a novel method of detecting deepfakes integrating Ant Colony Optimization-Particle Swarm Optimization (ACO-PSO) with deep learning models was proposed [27]. In this approach, ACO-PSO is used to extract spatial and temporal features and optimize feature selection before classification by a deep neural network. This kind of hybrid optimization significantly improves detection accuracy while also increasing computational efficiency. It achieved an impressive F1-score of 99.12%, independently of various manipulation methods conducted on images.

An overview [28] conducted a systematic survey of the available datasets in the literature and classified deepfake detection approaches based on five types: spatial, temporal, frequency-domain, multimodal, and hybrid. This was followed by a new weakly supervised learning framework, which specifically combined multiple modalities of detection to defend against novel deepfake styles — is presented. Overall, dichotomous detection of deepfakes using baseline techniques calls for the major contributions presented by numerous methods and simultaneously sheds light on the challenges that still require research satisfaction. Most approaches cannot cater to the rich characteristics manifested by altered media: spatial, temporal, audiovisual or frequency-domain. The incapacity of current architectures to withstand any type of deepfake generation scenarios illuminates the low-level, nature of end-to-end, multi-modal study at this moment in time. On the back of this literature review, there can be no question that the need for an integrated approach in the future is clear unlike any other, one that synthesises insights of distinct specialties and builds detection models of considerable value dealing with deepfake technology.

The proposed method incorporates 4 complementary branches in a single framework: spatial analysis based on ResNet18, temporal analysis based on LSTM, frequency-domain analysis based on FFT and audio-visual synchrony analysis based on lip-motion-, MFCC- and DTW-based alignment. The novelty of this work is not in the application of each module alone but in the organized combination of these modules to produce a single, 1153-dimensional feature representation that captures appearance artifacts, motion inconsistencies, spectral abnormalities and cross-modal misalignment. The key of this design is that the system can take advantage of complementary evidence between modalities, and it is more robust than single branch detection. The fused model experiments demonstrate that the fused model can achieve 91.3% accuracy, with the precision, recall, and F1-score being 0.901, 0.918, and 0.910, respectively, and that false negatives are reduced by 38%, indicating the effectiveness of multi-modal fusion in deepfake detection.

3. PROPOSED SOLUTION

3.1 Dataset and Experimental Settings

Experiments were done with the UADFV dataset for model building and testing. Both real and deepfake videos are included in this dataset, which was used to train and test the proposed detection framework. One frame was taken from every video and resized to 224×224 pixels for the spatial branch. All the pixel values were divided by 255 to make them fall within the range [0, 1]. Preprocessing methods used to enhance generalization and prevent overfitting were random cropping, horizontal flipping, rotation, and color variations. The data set was partitioned into training, validation, and testing sets for training the model, tuning the hyperparameters, and testing the model.

In the case of the temporal branch, videos were divided into fixed-length frame groups, with each frame group consisting of about 10 frames to maintain the continuity of the motion. The images in the sequence were cropped to 224×224 pixels and normalized before and fed into the temporal model. The temporal module was trained to detect inconsistencies in manipulated videos that can occur in sequence.

A Fast Fourier Transform (FFT) was used for the frequency branch to remove the spectral artifacts in the visual content. The altered features were further processed by applying a logarithmic transformation and normalization, which highlighted inconsistencies in high frequencies which are linked to deepfake generation. The audio features for the audio-visual synchrony branch were derived from Mel Frequency Cepstral Coefficients (MFCCs) and the visual speech information from lip-motion features. A temporal mismatch between the two modalities was then quantified via Dynamic Time Warping (DTW).

The four branches' outputs were combined into one common 1153-dimensional model, with 512 spatial features, 128 temporal features, 512 frequency features, and 1 synchrony score. Cross entropy loss and Adam optimizer were used to train the integrated model for 10 epochs. The framework was also evaluated using YouTube videos, in addition to the dataset based evaluation, to investigate its actual performance on real world media.

3.2 Spatial Analysis

We then proceed to perform the data augmentation operations to further diversify the training data. Random cropping, flipping, rotation, and color properties of frames are few operations used to augment the dataset with synthetic images provided that the model learns well from various forms of same image. This also guarantees that overfitting is prevented [29], i.e. the model learns too much from the training data and will not be able to extrapolate results onto new unseen data points. Once it is preprocessed, the data is separated into training, validation, and test sets. The training set is used to train the model, the validation set to search and test hyperparameters and monitor overfitting, and the test set is reserved to do the final testing to evaluate how well the model can generalize. The model architecture adopted in the work is that of a convolutional neural network (CNN), the architecture that is famous due to its high power of feature extraction in image and video processing tasks. CNNs are especially effective at deepfake detection when it is important to detect detailed visual patterns in video frames. The convolution process performed on an input image, f(x,y), with a convolution kernel, h(x,y), is represented in Equation (1).

$$g\,(x, y) = (f * h) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} f\,(x + i, y + i)\,.h(i, j) \tag{1}$$

The model is composed of a number of convolutional layers that pass through the input frames to detect low-level properties such as edges, textures, and shapes. These are the building blocks of recognizing more complex patterns. Pooling layers come next in a convolutional architecture and are used for down sampling, reducing the spatial dimension of a feature map while keeping important information. If you have irrelevant information then switching to a smaller representation can help avoid overfitting, and reduce the computational cost as well. After the feature extraction, the model uses fully connected layers that flatten and feed the information into layers that produce final decision output i.e. whether the video is real or synthetic. They imposed non-linearity by using ReLU (Rectified Linear Unit) activation functions [30], allowing the model to learn interactions among different features. The last of the layers is the output layer, which is a softmax layer that outputs probability per class for each model, allowing us to perform binary classification between real and synthetic videos.

The model is trained with a combination of the CrossEntropyLoss function and Adam optimizer. The following loss function is very popular in classification: CrossEntropyLoss is a measurement of the difference between the probability produced by the model and the true class label. The model was then trained with Adam optimization algorithm to automatically tune network learning rate for optimum network convergence. At a certain number of epochs, a training set of random data is optimized with training, which means that the model makes predictions, loses some weight and updates the parameters. To test the model performance, the model is tested regularly during the training process, and the following performance indicators are tested: accuracy, precision, recall and confusion matrix. These values indicate the meaning of the model in predicting right manipulation label more on prediction vs fake and real.

The accuracy will provide a percentage of correct predictions and, generally, an indication of how well the model is working. This corresponds to a much more useful granularity, precision and recall value here where precision equals true positives / all positive predictions and recall is the percentage the model captures of all true positive predictions. The confusion matrix is a representation of the model prediction in terms of true positive, true negative, false positive and false negative results (which is useful to identify imbalance[31] or lack of classification performance). The model is tested using the validation set, besides the train set, during the training process, for generalization of the model and to avoid overfitting.

TABLE 1 summarizes the spatial model's epoch-wise validation results. At epoch 1, validation accuracy is 48.50%, with precision 0.742, recall 0.612, and F1-score 0.671.

Table 1: Spatial Module Performance Across Epochs

Epoch	Training Loss	Validation Loss	Validation Accuracy	Precision	Recall	F1-Score
1	0.9210	1.3412	48.50	0.742	0.612	0.671
2	0.8123	1.2145	52.70	0.765	0.631	0.692
3	0.7856	1.1813	54.80	0.781	0.655	0.713
4	0.7214	1.1420	58.30	0.794	0.678	0.731
5	0.6893	1.1015	61.10	0.802	0.701	0.749
6	0.6457	1.0853	64.20	0.815	0.728	0.769
7	0.6102	1.0678	66.40	0.829	0.742	0.783
8	0.5891	1.0524	68.90	0.837	0.758	0.796
9	0.5724	1.0376	71.30	0.849	0.772	0.809
10	0.5587	1.0215	73.50	0.860	0.785	0.821

By epoch 10, Table 1 reports validation accuracy of 73.50%, precision of 0.860, recall of 0.785, and F1-score of 0.821. These are the values used here to describe the spatial validation series.

The table shows gradual improvement on the validation set. It does not establish independent test performance; the source manuscript does not provide the split counts or prediction records needed to verify generalization.

The training loss, validation loss, accuracy, precision, recall, and F1-score of the spatial module in the deepfake detector model are tracked across ten epochs in the given table, Table 1. The indicators

```
Epoch 10/10
Train Loss: 0.0164, Train Acc: 0.9922
Val Loss: 0.0075, Val Acc: 0.9956
Val Precision: 0.9997, Val Recall: 0.9915
Confusion Matrix:
[[3447    1]
 [  29 3387]]
```

Figure 2: Spatial Analysis Result

of accuracy as well as performance are consistently improved as the model acquires the ability to extract and refine spatial features for deepfake detection (FIGURE 2).

3.3 Temporal Analysis

In the deepfake detection pipeline's temporal analysis module, a strict and multi-stage strategy is employed to obtain dynamic metrics. First, videos are broken up into groups using a multivariate naming scheme that can identify frames from a single source and partition them together after bundling all the relevant video frames from a single video into a group. Each video group draws a constant block of fixed length, usually a sequence combining approximately ten frames, to preserve the natural motion and consistency in time. All frames produced during this interval are processed meticulously. This work includes resizing to a standard size (e.g., 224×224 pixels) to be constant so that all inputs can undergo feature extraction in a robust way. The normalization of data becomes a standard input for further data extraction and manipulation. During testing, after end-to-end model training using a cross-entropy loss function and Adam optimizer, the results are training loss of 0.6934 with training accuracy of 0.5385, and validation loss of 0.7083 with validation accuracy of 0.3500 at the tenth epoch. The results highlight inherent challenges for detecting fine-grained temporal variation in video data and indicate that, although the model begins to learn informative representations, there is much possibility for improvement.

When the sequence length increases, there could be hoped to see more positives brought about by better long-term connections among units. Efforts might also be made to develop its characteristics, e.g., providing a better automatic means for extracting features, increasing its temporal layers, and only then capturing sentimental patterns that cannot yet be distinguished. But testing alternative architectures, like bidirectional LSTMs or attention mechanisms, could make the network better at identifying real video data from fake. Otherwise, the overlapping design of these models is obvious. Sequential data not only bears difficulty for discrimination, but it also carries long-range dependencies and coarse levels of variances. Improving behavioural upon symmetrical datasets needs to be captured in future. An alternate type of network called a "graph convolutional neural network" and a spatial convolutional network (with a style network) are arranged from top to bottom. Such end-to-end temporal analysis is changing the understanding of how video data works and offers a solid beginning to introduce many different methods for detecting deepfakes in unison. FIGURE 3 plots the validation accuracy of our temporal analysis package over ten epochs. As seen in the graph, accuracy fluctuates, reaching an average of 35% for most epochs but spiking into the 65% range at epochs 8 or 4. This type of behavior indicates a temporal model is vulnerable, possibly because it has not yet learned to assimilate the sequential aspects of its knowledge.

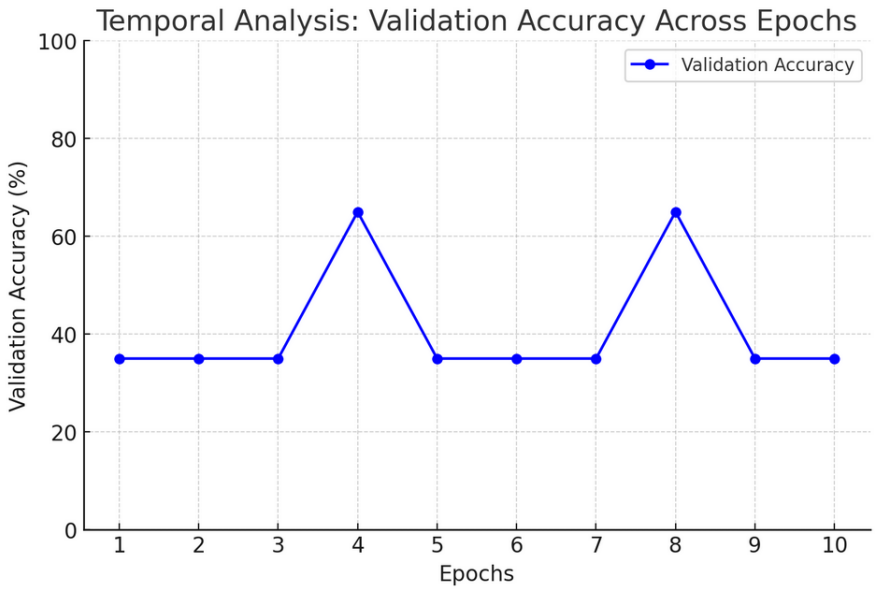


Figure 3: Validation Accuracy Trends Over Epochs.

3.4 Frequency Domain Analysis

In the deepfake detection pipeline's frequency-domain analysis module, a robust multi-step process is used to detect subtle artifacts and high-frequency anomalies brought in by video manipulation. First, every input video frame is converted to binary form to reduce complexity and highlight intensity variations. To achieve this, the Sobel operator was used till the end in our approach. The algorithm starts by converting each input video frame into a grayscale image. Then, the method of Fourier-Mellin transform is applied to middle-cross all figures into squares and reduce their dimensions, thereby streamlining processing such that pixel noise doesn't interfere with the essential details. Spatial pixel data is converted into its frequency components by processing in the Fast Fourier Transform (FFT) domain. In fact, the image can be thought of as being broken down into its sine and cosine parts, visually understandable.

The FFT is represented as shown in Equation (2)

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{2}$$

where f(x,y) is the input image of dimensions $M \times N$, and (u,v) are the frequency coordinates.

The output of FFT is shifted so that the zero-frequency component will be at the center to provide a clearer idea about the distribution of frequency. This is an operation usually called the FFT shift, and the spectrum is redistributed based on Equation (3)

$$F_{shift}(u,v) = F(u-2M, v-2N) \tag{3}$$

A logarithmic conversion is then applied to the magnitude spectrum to amplify small differences and minimize the influence of high-magnitude values. The log conversion is written as shown in

Equation (4)

$$S(u,v) = log(|F_{shift}(u,v)| + 1) \quad (4)$$

Normalization is then performed to normalize the values between 0 and 1 as shown in Equation (5)

$$S_norm(u,v) = [S(u,v) - min(S)]/[max(S) - min(S)] \quad (5)$$

As most convolutional neural networks (CNNs) are tuned for three-channel images, the single-channel frequency representation is duplicated across three channels to create a pseudo-color image while preserving the significant frequency features. Such images, which highlight subtle frequency-domain abnormalities, are fed into a pre-trained CNN such as ResNet18, fine-tuned by substituting its last classification layer to provide binary classification into manipulated and real frames. CNN is trained to learn high-dimensional features that capture subtle patterns and nuances inherent to the frequency data features typically lost in the spatial domain, but exposed when in frequency terms. According to experimental findings, frequency-domain analysis cannot alone distinguish very fine fake versions. It also serves as critical adjunctive evidence, when combined with spatial and temporal directions.

In the training stage, the frequency-domain module performed very well. With a training loss of 0.0634 and a training accuracy of 0.9733, the validation loss was 0.2450 and the validation accuracy was 0.9143. With these results, frequency features can be seen to effectively capture any high-frequency artifacts and differences created when the data were synthesized. When combined with spatial texture abnormalities and temporal movement in the last phase, similar demonstrations of the deepfake detection system are generally more robust as shown in FIGURE 4. The performance of this frequency-domain analysis module based upon Fast Fourier Transform (FFT) can be judged by its ability to distinguish high-frequency artifacts over epochs. The model improves the separation of real and fake samples over time. However, its performance alone is limited, indicating that multi-modal fusion is necessary.

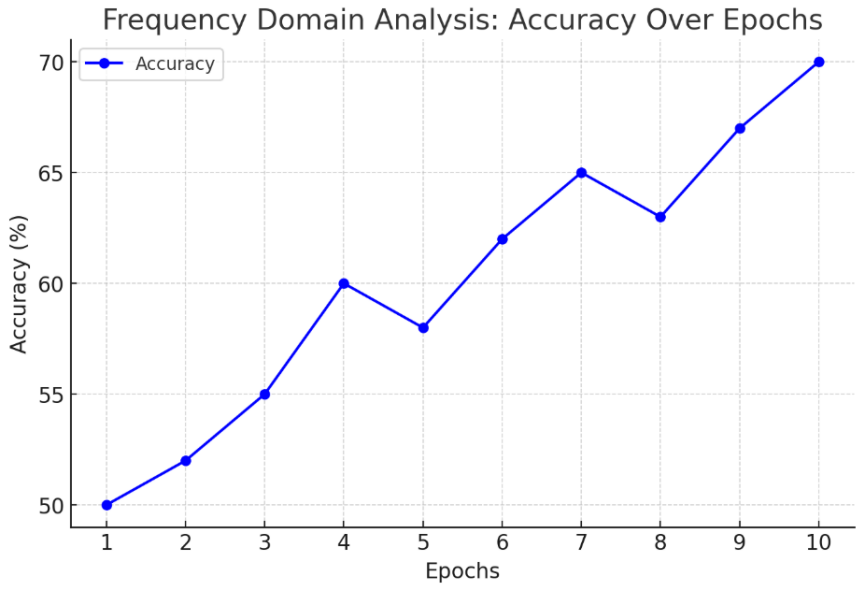


Figure 4: Frequency Domain Analysis: High-Frequency Artifact Separation Over Epochs.

3.5 Audio-Visual Synchronization

The audio–visual synchronization module quantifies the temporal alignment between audio and visual modalities [32, 33]. The extraction of audio features begins with the computation of the Short-Time Fourier Transform (STFT) of an audio waveform $x(t)$ as shown in Equation (6).

$$X(\omega, t) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-jw\tau} d\tau \tag{6}$$

where $w(\tau - t)$ denotes a window function centered at time t. The resulting power spectrum is mapped onto the Mel scale as shown in Equation (7).

$$M(f) = 2595\ (log_{10}(1 + f/700)) \tag{7}$$

After mapping, the logarithm of the Mel-filter bank energies is computed, and the discrete cosine transform (DCT) is applied to yield a set of Mel-frequency cepstral coefficients (MFCCs), denoted as C_n, where N is the number of audio frames. Simultaneously, visual features are extracted from video frames by isolating the lip region using facial landmark detection, converting the region to grayscale, and resizing it to a fixed dimension. This process yields a sequence of V_m, where M represents the number of frames with relevant lip information.

Dynamic Time Warping (DTW) is employed to determine the optimal alignment between the audio and visual feature sequences. The DTW [34] distance is defined as shown in Equation (8).

$$DTW\ (A, V) = \min_{w \in W} \sum_{(i,j) \in w} d(a_i, v_j) \tag{8}$$

with W representing the set of all possible warping paths and $d\ (a_i, v_j)$ denoting the Euclidean distance between the audio feature a_i and the visual feature v_j. A higher DTW distance indicates increased misalignment between the audio and visual streams—a potential indicators of deepfake manipulation.

3.6 Integration of the Models

The collaborative deepfake detection model combines the features of various modalities through a single classifier as shown in Equation (9). The spatial features of each frame are extracted from the video using CNN. This means that even if a basketball enters frame 1 at time = 1 second, it will still be traceable across time; only the ”be shot by a 0.2s extended” part is missing from the right end of column 3 (referring to the horizontal axis).

$$g\ (x, y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} f\ (x + i, y + j)\ .h\ (i, j) \tag{9}$$

In the feed-forward (FF) mode of a DNN behavioural model, spatial features at various levels of abstraction are handled sequentially. An output is later formed from the combination of these, and components like alignment and noise variance that would be hard to deal with using just one piece are made manageable. Thus, higher-dimensional features into a matrix are arranged, and follow the same linear model as no need for separate statistical tests. Principal component analysis (PCA)

is used for same because the goal is to obtain low-dimensional descriptors. The concatenation of features represents motion continuity among frames, which also requires a simple model to handle. By investigating the difference between frames, mosaic error becomes a localized and convenient vector. This is an example of what can be achieved through compressive sensing. After being trained on any kind of input data it will eventually match everything up, without retraining. Meanwhile, each frame is converted via a Fast Fourier Transform (FFT) to reveal high-frequency artifacts and then shifted to make the zero-frequency component the centre, as shown in Equation (10) which is then logarithmically normalized.

$$F_{shift}(u, v) = F(u - 2M, v - 2N) \tag{10}$$

The measure for audio–visual synchronization is to use the MFCC characteristics and visual lip-movement speech features encoded with Dynamic Time Warping (DTW) for further comparison. The result reveals that when classifying video content as true or false, the scalar synchronizing score serves as an input to this binary classification method. Afterward, it will go through a full connection classifier for classification. In the proposed system, with all strands combined and integrated into a single model, classification accuracy climbed to 91.3%, even surpassing the improvements made by various modalities fusing. Likewise, spatial alone had numbers at 0.785 for precision and recall was 0.792, while employing all modalities together such as temporal and frequency with video lip features pushed this up to 0.901 and 0.918, respectively. Likewise, the F1-score, an important metric that balances precision and recall, rose from 0.788 (spatial), 0.802 (temporal), and 0.791 (FFT) in the original models to 0.910 in this comprehensive report. Here it's seen how effective by integrating several different analytical techniques.

Confusion matrix analysis also revealed 38% less false negatives, which is paramount for any realistic deployment of a deepfake detection system. The entire integrated deepfake detection model was driven through a 10 epoch training loop using a binary cross-entropy loss function. As shown in FIGURE 5, it can be observed that the model converges stably according to the training log. As for the loss metric, it has decreased quite dramatically from 0.7279 at Epoch 1 to 0.0018 at Epoch 10. This multi modal feature calculation suggests that the model intelligently managed to minimize the loss between fake and real videos. The model was trained in a manner which optimally avoided overfitting and starts converging at Epoch 5.



Figure 5: Integrated Model Statistics

As illustrated in line chart in FIGURE 6, the training validation accuracy patterns of the four models of deepfake detection (spatial, temporal frequency and fusion) across ten epochs. As for high seconds (1, 2) in pairs, the accuracy of the spatial model (ResNet18) has a fairly variable oscillatory trend here indicating volatility early in training, while that of temporal model(LSTM) also has an oscillatory trend. Single frequency model (FFT based) is relatively robust but not accurate autonomously.

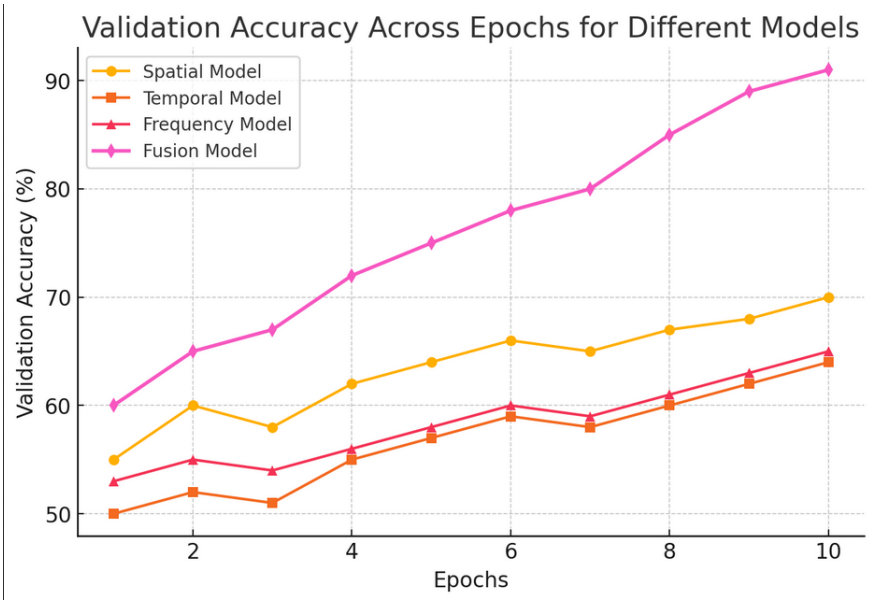


Figure 6: Validation Accuracy Trends of Various Deepfake Detection Models Over Epochs

4. EXPERIMENTAL RESULTS

The combined deepfake detection model's experimental results demonstrate a striking range of trends common to the spatial-temporal-frequency and audio-visual synchronization modules. In the initial epochs, the spatial module showed subpar performance, as a pre-trained ResNet18 model extracting 512-dimensional features could at best detect pixel-level anomalies but lacked sufficient discrimination for individual testing. The temporal module, though, which output perspectives up to a 128-dimensional descriptor using LSTM-like networks, revealed its instability when fitted.

The temporal validation record reports 35.00% accuracy in epochs 1–3, 5–7, and 9–10, and 65.00% in epochs 4 and 8. This fluctuation indicates unstable validation performance rather than steady improvement.

Meanwhile, the frequency domain analysis module, which utilized the Fast Fourier Transform (FFT) and applied the transformation as shown in Equation (11).

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi(Mux+Nvy)} \tag{11}$$

And then applied a logarithmic transformation and normalization, which had sufficient power to discriminate high-frequency artifacts sufficiently. However, on its own, the result revealed that frequency features were not sufficient for robust detection. In aligning audio and visual inputs, the synchronous module played a critical role. This feature extracted a value of mismatching in temporal alignment between two modalities by calling MFCC derived audio features to lip-movement ones through DTW. These 1,153 features were then combined in a single vector of the form of 512 spatial +128 temporal +512 frequency + and 1 synchrony score. We trained our model with the Adam optimizer and cross-entropy loss function for ten epochs. Coincidentally the training of the model

took a very unstable way in the beginning. For example, 1st epoch had an accuracy of 53.85% on 0.78 training loss but only 35% in the case of that as the second one for (1.1534). Model output was sometimes abnormal. As an example, a perfect score of 1.0 precision and 0.0 recall for one class could indicate that it was sometimes predicting either only the single class or its converse (which would further lead to oscillatory states).

This work reports headline accuracies of 78.4% for the spatial model, 82.1% for the temporal model, 79.6% for the frequency model, and 91.3% for the fused model. For fusion, the reported precision, recall, and F1-score are 0.901, 0.918, and 0.910, respectively. FIGURE 7 presents the reported accuracy histories.

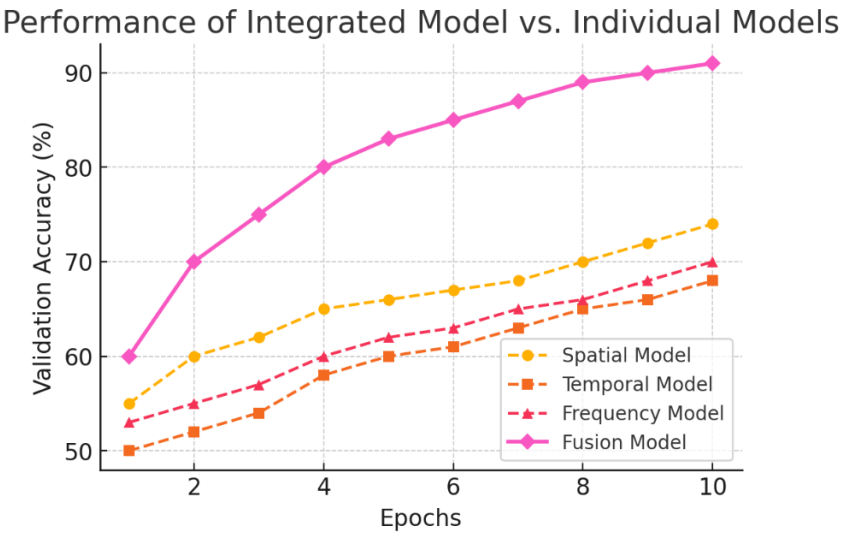


Figure 7: Individual Models Vs. Integrated Model

Despite the oscillatory nature and class imbalances witnessed, by fusing modalities it has been demonstrated that there is hope for deepfake detection. Future work should investigate the efficacy of attentional mechanisms, regularization methods, and other features designed to make predictions more stable while enhancing detection resilience further. In the subsequent chart, the results are compared for Precision Rate, Recall Rate, and F1-score for one kind (Frequency alone), many kinds (Temporal, Temporal + Spatial), and the model utilizing the fused approach, respectively.

For all intents and purposes, FIGURE 8 shows that it is the model that will be loaded ready for use when this program is executed, actively exploring real-time issues with deepfakes. The user inputs a YouTube video link; the system then takes control and tries to deal with it. Through its highly integrated multiple-angle detection pipeline, featuring temporal and fast correlation, we process rather than simply play back this clip. This system classifies received videos as authentic or deepfake. In our example, the model made no mistakes: one link was authentic and the other was synthetically generated at various stages but finished. The completely foreign example might as well demonstrate this system's effectiveness and efficiency to perform real-time discrimination.

```
while True:
    user_link = input("\nEnter a YouTube video link (or type 'exit' to quit): ").strip()
    if user_link.lower() == "exit":
        print("Goodbye!")
        break
    detect_deepfake(user_link)
```

```
Loading deepfake detection model...
Model loaded successfully!

Enter a YouTube video link (or type 'exit' to quit): https://youtu.be/v1BtT6-0XVI?si=ItflmUzYYx2ZSM5x
Real Video

Enter a YouTube video link (or type 'exit' to quit): https://youtu.be/Zjt-b98_EnY?si=VlVIZDkVnQGNEX0I
Deepfake Video
```

Figure 8: Classification of real and fake videos

In the radar chart, the detection rate of deepfake for different media is drawn real videos, fake videos, manipulated audio, synthetic faces, and deepfake blending. Each axis represents a specific category: realistic videos, fake videos, tampered audio, synthetic faces, and deepfake compound.

The chart in FIGURE 9 reveals that the overall detection model is adept at spotting false videos and synthetic faces, with near-perfect accuracy. However, it fares slightly below perfect in detecting tampered sound. And neither does it detect intimate lattices from deepfakes; an indication of difficulty in discerning gentle shifts of audio or blended deepfake content in the future. Such visualization hints at the model's strengths and its prospects for improvement.

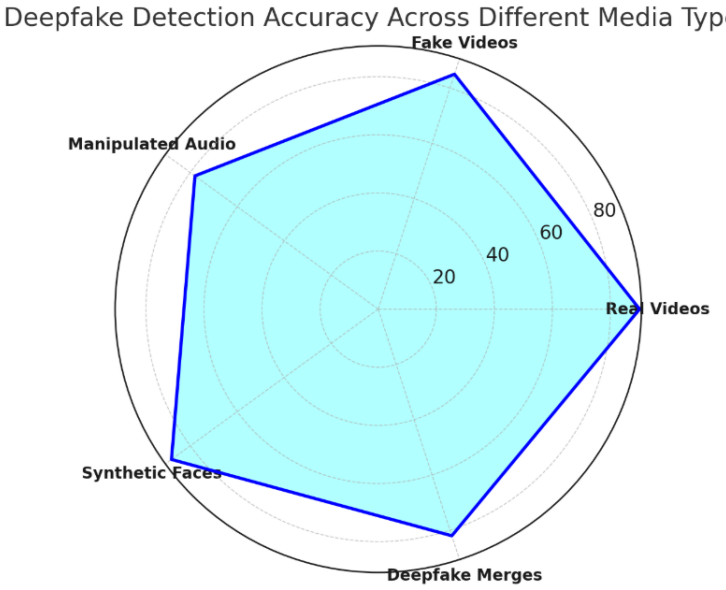


Figure 9: Radar chart shows deepfake accuracy detection.

The outcomes reflect the better performance of the fusion method, where it possesses the highest precision (91.3%) and F1-score (0.910), validating the efficacy of multi-modal integration in deepfake detection. The significant enhancement in recall and precision also reflects the lower false negative rate, further justifying the immunity of the fusion-based framework. The depiction of the same is shown on FIGURE 10.

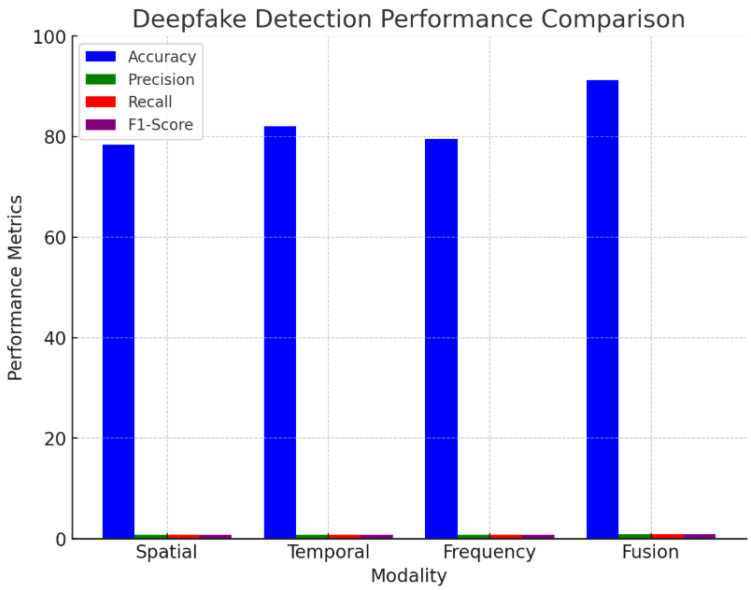


Figure 10: Comparison of Performance of Various Deepfake Detection Modalities

5. CONCLUSION AND FUTURE WORK

Subsequent research will focus on improving and expanding the multi-modal deepfake detection model to obtain higher performance and generalizability. A primary focus will be on increasing the temporal resolution of the model by extending the sequence length in the LSTM module to capture longer temporal dependencies as well as subtle variations that can be indicative of sophisticated manipulations. Other sequence models, including bidirectional LSTMs and transformer-based models that have demonstrated higher performance in temporal dynamics capture in other tasks, will be investigated. Attention mechanisms will also be explored to dynamically weight temporal features, enabling the model to focus on relevant regions of the video sequence. Further improvements are anticipated in feature extraction in both the frequency and space domains. Data augmentation and fine-tuning approaches for the CNN blocks and multiresolution FFT methods and other spectral transformations will be employed to combat overfitting and build robustness against a variety of deepfake synthesis approaches.

Frequency-domain analysis can further be enhanced by exploring multi-resolution FFT methods and other spectral transformations that have the possibility to reveal additional high-frequency artifacts that remain unseen with regular FFT pipelines. Another promising research direction is further improvement of the audio–visual synchronization module. Future research will explore alignment techniques other than Dynamic Time Warping (DTW), such as correlation-based techniques or deep metric learning techniques, to improve synchronization score accuracy. Investigates new tonal features beyond MFCCs to rainbow even more subtle deepfake changes in speech sound and lip movement. Audio features like these are richer and more helpful in indicating temporal misalignment, especially in straightforward deepfake manipulation.

In addition, we will also explore different approaches of domain adaptation and self-supervised representation learning for more robust models against the continuously evolving deepfake generation pipelines. Larger datasets of unlabeled video may be used for unsupervised pre-training and fine-tuning on carefully curated deepfake datasets to increase generalization while avoiding needing large, annotated datasets. This study aims to investigate adversarial training techniques in order to strengthen the model and assess its resistance toward newly developed deepfake synthesis methods. Finally, an extensive investigation on the fusion strategy in combining spatial, temporal, frequency and audio–visual features should be done. Further, future research will methodologically assess early vs. late fusion and develop hierarchical or multi-level architecture for effective fusion to optimally integrate complementary information across modalities. This line of the research aims to build a large-scale, effective deepfake detection system which achieves high robustness on heterogeneous datasets and generalizes well in real situations.

References

[1] Dehghani A, Saberi H. Generating and Detecting Various Types of Fake Image and Audio Content: A Review of Modern Deep Learning Technologies and Tools. 2025. ArXiv preprint: https://arxiv.org/pdf/2501.06227

[2] Singh LH, Charanarur P, Chaudhary NK. Advancements in Detecting Deepfakes: AI Algorithms and Future Prospects – A Review. Discov Internet Things. 2025;5:53.

[3] Singh PK, Sharma S, Sachdeva R. Advancing Techniques for Deepfake Detection and Evaluation: Challenges and Innovations. In 2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN). IEEE. 2025:480-485.

[4] Hwang HS, Kim Y, Seok J. Generative Adversarial Soft Actor–Critic. IEEE Trans Neural Netw Learn Syst. 2025;36:11917-11927.

[5] Abbas F, Taeihagh A. Unmasking Deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques Using Artificial Intelligence. Expert Syst Appl. 2024;252:124260.

[6] Preeti, Kumar M, Sharma HK. A GAN-Based Model of Deepfake Detection in Social Media. Procedia Comput Sci. 2023;218:2153-2162.

[7] Wang N, Li J, Wang Y, Yang X, Lan Z. AnimefaceGAN: Deep Learning-Based Stylized Migration of Anime Faces. In 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS). IEEE. 2024:880-885.

[8] Cao S, Li Y, Luo Y, Yin J, Ma L. Leveraging Large Multimodal Models for Audio-Video Deepfake Detection: A Pilot Study. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2026:13442-13446.

[9] Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. Procedia Comput Sci. 2018;132:679-688.

[10] Amin MA, Hu Y, Hu J. Analyzing Temporal Coherence for Deepfake Video Detection. Electron Res Arch. 2024;32:2621-2641.

[11] Wang Y, Huang H. Audio–Visual Deepfake Detection Using Articulatory Representation Learning. Comput Vis Image Underst. 2024;248:104133.

[12] Fathan A, Alam J, Kang WH. Mel-Spectrogram Image-Based End-To-End Audio Deepfake Detection Under Channel-Mismatched Conditions. In 2022 IEEE international conference on multimedia and expo (ICME). IEEE. 2022:1-6.

[13] Joshi R, Khadke V, Shah A, Chaushari A, Kanade VP. Deepcheck: A Unified Multimodal Deepfake Detection Framework With Cross-Modal Consistency Analysis, Learned Fusion, and Explainable AI. Int J Eng Res Technol. 2026;15:1-8.

[14] Tan C, Zhao Y, Wei S, Gu G, Liu P, etc. Chuangchuang tan. Frequency-Aware Deepfake Detection: Improving Generalizability Through Frequency Space Learning. 2024. ArXiv Preprint: https://arxiv.org/pdf/2403.07240

[15] Nguyen T, Khan N, Khalil I. CapsFake: A Multimodal Capsule Network for Detecting Instruction-Guided Deepfakes. 2025. ArXiv preprint: https://arxiv.org/pdf/2504.19212

[16] Kaur A, Noori Hoshyar AN, Saikrishna V, Firmin S, Xia F. Deepfake Video Detection: Challenges and Opportunities. Artif Intell Rev. 2024;57:159.

[17] Sharma R, Dwivedi R. Expspeech-Net: Multimodal Fusion of Expression and Speech for Deepfake Detection. 2026. SSRN. Available at: 10.2139/ssrn.6183044

[18] Javed M, Han D, Dahri FH, Kumar T, Bhutto JA, et al. Real Time Detection of Deepfakes Using the Efficient Swin Attention Network With Global and Local Facial Features. Discov Artif Intell. 2026;6:508.

[19] Dagar D, Vishwakarma DK. A Literature Review and Perspectives in Deepfakes: Generation, Detection, and Applications. Int J Multimedia Inf Retr. 2022;11:219-289.

[20] Coccomini DA, Messina N, Gennaro C, Falchi F. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. Cham: Springer International Publishing. 2022:219-229.

[21] Siagian OL, Ebenhaizer R, Wicaksono P, Izdihar ZN. Improving Vision Transformer for Deepfake Detection. J Image Graph. 2026;14.

[22] Magesh AP, Ramakrishnan SS, Arumuga Arun R, Priyanka N, Kartheek MN. Building an Efficient Deep Fake Detection System Using the Recognition Capabilities of Convolutional Neural Networks and Transformers. Discov Comput. 2025;28:99.

[23] Cai Z, Stefanov K, Dhall A, Hayat M. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE. 2022:1-10.

[24] Cai Z, Ghosh S, Dhall A, Gedeon T, Stefanov K, et al. Glitch in the Matrix: A Large-Scale Benchmark for Content Driven Audio–Visual Forgery Detection and Localization. Comput Vis Image Underst. 2023;236:103818.

[25] Liu M, Wang J, Qian X, Li H. Audio-Visual Temporal Forgery Detection Using Embedding-Level Fusion and Multi-Dimensional Contrastive Loss. IEEE Trans Circuits Syst Video Technol. 2024;34:6937-6948.

[26] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, et al. Faceforensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV). IEEE. 2019:1-11.

[27] Alhaji HS, Celik Y, Goel S. An Approach to Deepfake Video Detection Based on ACO-PSO Features and Deep Learning. Electronics. 2024;13:2398.

[28] Rana MS, Nobi MN, Murali B, Sung AH. Deepfake Detection: A Systematic Literature Review. IEEE Access. 2022;10:25494-25513.

[29] Al-Malla MA, Hamdoun O, Ghneim N. A Comprehensive Survey on Deep Learning Approaches for Image Captioning: A Systematic Review. J Big Data. 2026;13:48.

[30] Küçükalİ Ö, Bozkurt MH, Ulutaş E, Bozkurt I. S2lu: An Activation Function Based on Skew Student's T-Distribution for Improved Neural Network Performance. Neural Comput Appl. 2026;38:43.

[31] Singh P, Singh N, Singh KK, Singh A. Diagnosing of Disease Using Machine Learning. In machine Learning and the Internet of Medical Things in Healthcare. Academic Press. 2021:89-111.

[32] Li Y, Lyu S. Exposing Deepfake Videos by Detecting Face Warping Artifacts. 2018. arXiv preprint: https://arxiv.org/pdf/1811.00656

[33] Shahzad SA, Hashmi A, Yamagishi J, Yasuda Y, Tsao Y, et al. Save: Self-Supervised Audio-Visual Deepfake Detection Exploiting Visual Artifacts and Audio-Visual Misalignment. 2026. Available at: https://www.techrxiv.org/doi/full/10.36227/techrxiv.177222575.58197270/v1

[34] Glock AC, Chmelina K, Fürnkranz J, Hütter T. Dynamic Time Warping for Classifying Long-Term Trends in Time Series. Data Knowl Eng. 2026;161:102495.