International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

Deep Learning Framework for Detecting Deep Fake Media Using Lip Region Analysis and Audio–Visual Synchronization

Hitashri M [1], Deekshitha M [2], Boreddy Sahastra Reddy[3], Dr. Nagaraj Cholli[4]
[1,2,3]UG, Dept. of CSE, CMR University, Bangalore, Karnataka, India.
[4]Professor & Associate Dean, CMR University, Bangalore, Karnataka, India.
Email ID: *shrihi1204@gmail.com[1], deekshitha1325@gmail.com[2], sahastraboreddy@gmail.com[3] , drnagaraj.c@cmr.edu.in[4]*

Abstract

The advancement of the 'deep fake' technology has been rapid, with the emergence of innovative generative models that have the capability to create realistic fake media. Such fake videos have the potential to impact the authenticity, security, and trust of the information being conveyed, thereby creating a need for the detection of such fake videos. Several methods for the detection of 'deep fake' videos have been proposed, which primarily focus on the global facial artifacts and temporal inconsistencies. However, such methods may not perform well with the emergence of advanced 'deep fake' models, which have the potential to generate realistic fake videos. In this paper, we have proposed a 'deep fake' video detection framework that primarily focuses on the lip region and the audio-visual synchronization. The process was carried out through a series of checks, progressing through a series of steps in a specific order. The process starts by closely inspecting the video, followed by inspection of the lip region, inspection of any cues present in the video, and then audio features, among others, until a conclusion is reached. A major portion of this process is focused on the lip region, which is used to identify the speaker in a three-dimensional space. Audio cues, including Mel-frequency cepstral coefficients, are also considered during this process. The process involves authentic and manipulated videos, and it is clear that unusual lip movement is a strong indicator of a deep fake video.

Keywords: *Deep fake Detection, Audio Video Synchronization, Lip Region Analysis, Video Forensics, Artificial Intelligence, Multimedia Security.*

1. Introduction

Deepfakes are a type of AI tech used to make a person say something they never said. "Deepfake" is a combination of "Deep Learning," an important concept in artificial intelligence, and "Fake." Deepfakes were found in 2017 when a face-swapped video surfaced on the internet. Having the capabilities of the most potent artificial intelligence architectures, such as Generative Adversarial Networks (GANs), and the most recent forms of diffusion, deepfakes have the potential to merge fake content with real content in a seamless manner [6], with studies suggesting that only 50–60% of them can be detected by the average, non-trained individual. This emergence of deepfakes has led to some serious issues in various fields, including politics, where the election process is threatened by the emergence of inflammatory speeches such as videos of world leaders declaring wars, finance, where voice-based deepfakes are causing scams resulting in the loss of billions of dollars, and security, where the threat is used for the theft of identities and propaganda. The menace of deepfakes was first experienced in 2018 when a video was produced by BuzzFeed using AI, with the assistance of filmmaker Jordan Peele, in which the face and voice of the former U.S. President, Barack Obama, were imposed on the same physical body, stating false and alarming information about political occurrences. In the case of the Indian subcontinent, the issue came into play in the context of the 2023 Rashmika Mandanna deepfake, in which a video went viral featuring the actress's face on another woman's body in an elevator scene, garnering millions of views before the outrage, Delhi Police FIR, and call for stricter regulations on AI

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

came into play—the culprit behind the crime was caught, and the role of the medium in the crime of violating privacy and harassment was evident. There have also been several cases of fake videos involving billionaire Mukesh Ambani, such as the creation and circulation of fake videos involving him at the Vibrant Gujarat Summit promoting fake schemes of stock trading, resulting in financial scams and the resulting fraud on investors. With the availability of tools for creating deep fake media, e.g., through the availability of open-source models based on Stable Diffusion variants, the sheer volume of such media is creating issues, thereby exacerbating the threat to the stability of society at large. Conventional methods for authenticating media based on the overall inconsistency of the face and audio/video desynchronization have been observed to not perform well in the face of state-of-the-art deep fake media, thereby necessitating the need to analyze the specific features, e.g., the region around the lips, wherein the biomechanical inconsistency in movement, texture, and synchronization is likely to expose the fake nature of the media. The world is shifting at a swift rate to resolve the dilemma of deepfake with the aid of technology and regulations. The watermarking method, such as Google's SynthID and Adobe's Content Authenticity Initiative, is being utilized by governments and tech companies to ensure that AI-generated content is infused with invisible tags, and the use of blockchain is also being leveraged to track the source of the content in real time. On the contrary, tech companies such as Microsoft and Meta are utilizing deepfake classifiers with the aid of multimodal AI, enabling the detection of deepfakes with a high accuracy rate of 90% based on frequency, eye blink, and spectrum features [8]. However, there are certain problems that persist when it comes to picking out deepfakes. First, the detectors may not work for other deep fake generators if they were only trained on one. Secondly, they are very costly when we use it in use case. Moreover, the features of few other deep fake generators may affect the reliability. Which is why it has, led to the creation of many methods for identifying deepfakes. These methods utilize blood flow biometric signals based on rPPG and employ the MesoNet algorithm for anomaly detection in texture signals. In the present study, audio-visual signals are added using transformer-based methods like the AV Lip-sync model. However, it is observed that while making talking deepfakes, the lip area is not considered, even though it is related to the relationship between viseme and phoneme. The process works by using a simple step-by-step approach to process the video. The video is first processed to obtain frames. The frames are then used to obtain the visual features. The visual features are mainly obtained from the lip region since the lip region is an important region of the face. This is a general approach to deepfake video detection since the face movements are usually altered. The process is divided into two steps. The steps are to obtain the visual features and the audio features. The audio features are obtained by analyzing how well the video synchronizes with the lip movements. This is important since the phonemes and visemes are usually difficult to match when the video is a deepfake. The last step is to analyze how well the video synchronizes with the visual features. This information is sent to a decision module to determine whether the video is a deepfake or not. The decision module also provides a confidence level. The study intends to investigate whether deepfake videos can be identified using lip reading and audio-visual cues.

2. Related Work

Deepfake identification is a popular research topic due to the increased potential of the newly developed models. Various techniques for detecting manipulated media have been explored in recent research, including facial artifact analysis, temporal behavior modeling, physiological signal analysis, and audio-visual synchronization verification techniques. These techniques aim at detecting the inconsistencies present in manipulated media, which may arise during the artificial generation and processing of the video content. One of the initial works on audio-visual relationships for video analysis was presented by Chung and Zisserman [1], where they proposed an audio-visual synchronization framework using a two-stream convolutional neural network architecture. In this technique, the authors aim at learning the embedding between the audio and mouth movements using the video content, where short audio-visual

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

segments are processed and compared based on the synchronization between the audio and mouth movements. Using this technique, it is determined whether the audio and mouth movements are synchronized, thus providing cues for detecting manipulated media. Some studies have also been conducted on analyzing certain facial areas where the artifacts of manipulation are more likely to be visible. In one such study, Hari Prasad et al. [2] proposed a method for detecting deepfake videos by focusing on the lip area. In their proposed method, the video is segmented into frames, and lip segmentations are carried out using statistical methods such as the Minimum Covariance Determinant (MCD) method. By analyzing the lip movements in the video, the method tries to detect abnormal patterns, which are likely to be present in manipulated videos. This study demonstrates the importance of analyzing certain facial areas where abnormal patterns are more likely to be visible. Some researchers have also focused on analyzing the behavioral patterns of facial movements in detecting deepfake videos. Li et al. [3] proposed a method for detecting deepfake videos based on eye blinking patterns using convolutional neural networks and LSTM networks. Temporal modeling of the frames in a video sequence also found its place in the domain of deepfake detection. A method in this direction was proposed by Guera and Delp [4]. This method makes use of convolutional neural networks to carry out the task of spatial feature extraction and recurrent neural networks for temporal modeling. In a similar direction, Agarwal et al. [5] carried out a study on the detection of deepfake videos using facial appearance and behavioral inconsistencies. This study focused on the detection of deepfake videos using different facial characteristics such as head movements, facial expressions, and changes in appearance. This study concluded that the detection of deepfake videos can be carried out using a combination of different facial characteristics. According to the research above, in order to tell if a video is a deep fake, you need to look at a lot of frames over time and space. Of all the parts of the face that are checked for deep fake detection, the lips are the most important. Lip movement is very complicated when a person is talking. In a real video, the lips and the sound go pretty well together. A lot of researchers have also studied the skin, blinking, and other parts of the face, but the lips are the most important for deep fake detection. If a video is a deep fake, the lip movement and the sound don't go well together. When a person is talking, a lot of things are moving, especially the lips. In a real video, the lip movement and the sound go very well together. But in a deep fake, a little bit of a difference will occur, and this is because the sound and the video are recorded separately. Normally, the lip movement for a phoneme or a viseme should match, but this doesn't occur in a deep fake. So, this study is based on the lips, and the video sound is checked with the lip movement for deep fake detection. Table I shows the key features of the current research on deep fake detection. From the table, it is very clear that most researchers focus on the lips for deep fake detection

Table 1 Comparison of Existing Deepfake Detection Approaches

Study	Key Detection Strategy	Relevant Features Considered In This Study
Chung and Zisserman	Audio–visual synchronization using two-stream CNN architecture	Audio–visual synchronization between speech and lip movements
Hariprasad et al.	Lip-region anomaly detection using statistical methods	Lip-region motion analysis and localized facial feature examination
Li et al.	Temporal eye-blinking detection using CNN and LSTM models	Temporal analysis of facial movements across frames
Guera and Delp	Temporal frame analysis using CNN and RNN models	Sequential frame relationships and temporal feature modeling

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

Agarwal et al.	Facial appearance and behavioral inconsistency detection	Facial motion patterns and behavioral feature analysis

3. Proposed Methodology

The framework, which has been proposed, will be useful in detecting deepfake videos based on the inconsistencies in the lip movement and the audio. Currently, with the advancements in AI, there are many tools available which can be used to manipulate the lip movement in the video and match it with the audio. Although the video will look real, the lip movement and the audio will not be perfectly in sync. Generally, the speech of any person follows a certain pattern, and the lip movement will be in sync with the audio. But if the pattern seems to be off, then there will be inconsistencies in the lip movement. As Shown in Figure 1.

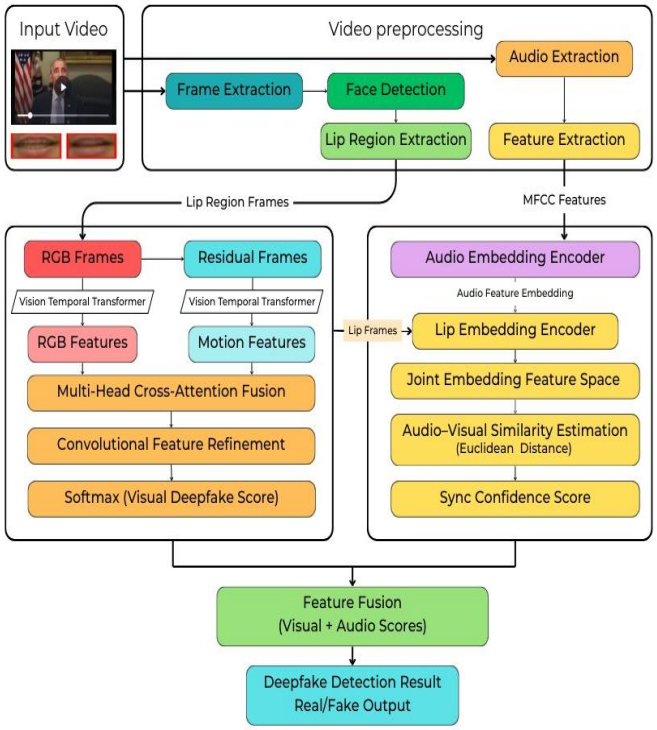

Figure 1 System Architecture for Audio-Visual Deepfake Detection

Figure 1 illustrates the overall architecture of the proposed system. The framework processes the input video through multiple stages, including video preprocessing, lip region extraction, visual feature analysis, audio feature extraction, audio–visual synchronization verification, and final classification. Each stage of the proposed system is described in detail in the following sections.

3.1 Video Preprocessing and Frame Extraction

The first step in this system is the preparation of the input video for further analysis. The input video is decoded and converted into a sequence of video frames while maintaining the order in which these video frames appear in the video. Let the input video be represented as:

$$V = \{F_1, F_2, F_3, \dots, F_n\}$$

where F_i is the i^{th} frame obtained from the video and n representing the total count of frames present. Each frame is subject to basic processing, where reconfiguring, normalization, and noise filtering are done. The reconfiguring ensures all frames are of the same size, which is helpful for calculations. Once all frames are extracted, the system will look for the person's face. Once the face is located, the system will look for the lips. Once the lips are located, the system will look for the part of the frame where the lips are most active. Finally, the system will separate audio and video.

3.2 Lip Area Split

After the face is located, the system then locates the lips. Facial landmarks are then used to locate the lips. If the detected facial region is represented as R_f, then the lip region can be expressed as:

$$R_l \subset R_f$$

where R_l represents the cropped mouth area extracted from the face. Lip region frames are used in the process. The aim of the process is to examine the lips. It is easier to understand the movement of the lips while a person is speaking by focusing on the lips. During the process, unwanted parts of the video are removed. They may include the background or the face.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

3.3 Lip Frames Learning

At this point, the speech features are tested. After the lip region frames have been extracted from the video, the lip visuals are examined while the person is speaking.

- **RGB Lip Frame Analysis:** In this method, the RGB lip frames are used to analyze the features. The Vision Temporal Transformer learns the features from these frames.
- **Residual Motion Analysis:** This method is referred to as the motion method. In this method, motion is depicted by residual frames. It represents motion from one lip frame to the next by examining the difference between frames, as shown in Equation 3.

$$D_t = F_{t+1} - F_t$$

In this equation, F_t represents the lip frame of the video at time t. This depiction of motion helps the vision temporal transformer learn motion features of lip movements during speech. These features will be able to detect unusual lip movements that may be indicative of deep fake videos. After learning these features using both methods, they are combined by using multi-head cross-attention. Then, convolutional layers and a softmax layer classify these features.

3.4 Audio Feature Extraction and Synchronization Analysis

Audio features are very important. Audio features can be defined by Mel-Frequency Cepstral Coefficients (MFCC). Let the audio waveform be represented as $A(t)$, where $A(t)$ represents the amplitude of the audio signal at time t. From this signal, MFCC features are extracted and represented as:

$$X_f = \{f_1, f_2, f_3, \ldots, f_q\}$$

Here, q represents the number of audio features that are extracted. Once this process is completed, the audio signal is passed through an audio embedding encoder that compresses it into a compact representation of the spoken content. At the same time, the extracted lip frames are input into a lip embedding encoder that generates corresponding lip embedding's. At this stage, it becomes possible to measure how similar or dissimilar the audio signal is with respect to the observed lip movements. This similarity is computed by calculating the Euclidean distance between the audio embedding's and lip embedding's. The distance can be expressed as:

$$d = || E_a - E_v ||_2$$

where E_a represents the audio embedding and E_v represents the lip embedding. When the computed distance is small, it indicates a high degree of similarity between the lip movements and the audio signal, which is expected in authentic videos. Conversely, a larger distance indicates a lack of synchronization between the audio and lip movements, which may suggest possible video manipulation. Based on this distance measure, a synchronization confidence score is obtained.

3.5 Feature Fusion and Final Deepfake Detection

At the final stage of the proposed pipeline, the information obtained from the visual analysis branch and the audio synchronization branch is integrated to make the final deepfake detection decision. Let the visual deepfake score be represented as S_v, and the synchronization confidence score be represented as S_a. The final decision score can be represented as:

$$S_{final} = f(S_v, S_a)$$

where $f(\cdot)$ represents the fusion function that combines the two scores. Based on this fused score, the system classifies the video as either real or manipulated:

$$y = \begin{cases} 0, & Real\ Video \\ 1, & Deepfake\ Video \end{cases}$$

The final output provided by the system consists of the classification outcome and a confidence measure that signifies the probability that the video is manipulated. It examines the movement of the lips and whether they match the sound in the video. This helps in identifying changes in deepfake videos. It

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

examines the sound in the video and how it relates to the movement of the lips. It then examines whether the sound in the video relates to the movement of the lips. This helps in identifying whether the video is real or not. The basic idea here is that real videos always have matching sounds and lip movements. However, deepfake videos lack matching sounds and lip movements. Deepfake videos cannot match this requirement.

4. Results and Analysis

After the process is done, the system can tell if the video is real or fake. Then the lips move with the sound. From the results, you can tell if the video is real or fake, the lips move with the sound, and the video looks real. First, the system will check if the video is real or fake. It will look at the video to tell. Then the lip area will be removed. After that, the RGB lip frame and the motion frame will be shown. When the lips move with the muscles underneath them, the lips are real. When the lips move with the muscles underneath them but not exactly, it may mean the best model did not copy the lips well. When the lip frame is ready, audio is processed too. This involves describing the audio and lips. This is done using Mel Frequency Cepstral Coefficients. Then audio is added to the lip frame using audio-visual sync. If the video is real, then the audio and video agree with each other. If the video is fake, then they do not agree with each other. If the lip frame with the RGB lip frame is real, then the lips may not move with the background. You may also not see a gap in the motion frames. But the system can only do so much. If you are shown two frames, then they may not have been made at the same time. Therefore, the sync score may not be high. Tests are carried out on real and fake videos. The tests show that the program works as it should. The one that gives two outputs also works as it should. In this program, both audio and video are analyzed. Synchronization scores are used. Therefore, a decision is made based on more than one type of data. This is different from the previous deep fake detectors. The tests show that you cannot be sure if a video is real or fake just based on the audio and video. What matters is if the audio and video agree or disagree with each other. As Shown in Figure 2.

Figure 2 Lip Region Extraction and Motion Representation in A Real Video (Kennedy Sample)

These visually extracted features are further used in the classification model and the synchronization module to ascertain the correspondence between the lip movements and the audio signal. For the second evaluated case, the system processes a modified video sample in which the lip movements are artificially generated in order to match a different audio signal. Similar to the previous case, the frames in the lip region are extracted and the motion representation is computed. However, in the residual frames, irregular motion patterns and pixel variation are often found due to the artificially generated lip movements, which do not mimic the natural patterns of human speech. As Shown in Figure 3.

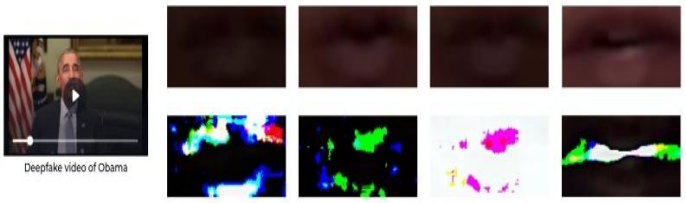

Figure 3 Lip Region Analysis of a Manipulated Video (Obama Deepfake Sample)

In the testing phase, a number of video samples were processed in order to evaluate the changes in the classification and synchronization values generated by the system. For one of the samples under consideration, the fake probability generated by the system was found to be 99.23%, whereas the actual probability was merely 0.77%. This revealed a high possibility that the video under consideration had undergone manipulation. For this particular sample, the system had generated a synchronization confidence score between the lip and audio signals as 0.32. This revealed a low synchronization between the two signals, and hence the system had classified

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

the video as a deepfake since a lip and audio mismatch had been detected. For the second evaluated video, the system reported a fake probability of 0.07%, but the actual probability reported by the system was 99.93%, which is very high and suggests that the video is authentic. In this study, the synchronization block reported a lip audio synchronization confidence score of 0.84, which implies a stronger synchronization between the audio and the video. The pipeline indicated a strong audio and lip synchronization, which is mostly expected for real videos. From the above analysis, it is observed that the synchronization confidence score reported by the pipeline is helpful in backing the classification outcomes. Videos with low synchronization confidence score values have greater fake probability score values, implying that the videos could have been altered. The above discussion suggests that the audio and video synchronization block helps enhance the accuracy of the detection process. The experimental outputs derived from the proposed system prove that the integration of visual feature analysis for the lip region and audio-visual synchronization evaluation can efficiently emphasize the discrepancies that are likely to occur in the manipulated media. The proposed framework for the detection of lip synchronization deepfakes is based on the analysis of motion patterns within the lip region and audio-visual synchronization. A summary of the observed results is presented in Table 2.

Table 2 Summary of Observed Results

Test Sample	Fake Prob. (%)	Real Prob. (%)	Lip–Audio Sync Confidence	System Observation
Sample 1	99.23	0.77	0.32	Lip–audio mismatch detected (classified as deepfake)
Sample 2	0.07	99.93	0.84	Strong lip–audio synchronization detected (classified as real)

Conclusion

The rapid development of fake media using deep learning makes it difficult to verify if a video is real. This study aims to verify if a video is real by checking if the lips and sound match. The process of video honesty check involves several steps. The results of this study show that if the sound does not match the lips, then the video is fake. However, if the sound matches the video, then the video is real. This is how deep learning is used to create fake media. The study of the lip area has several advantages. This study did not check the entire video. Instead, it only checked the lip area. This makes the video honesty check better than if the study checked the entire video. Another advantage of this study is that it checks how well the sound matches the video. The study used MFCC to check for video honesty. This makes the video honesty check work well to check if the sound matches the video. Another advantage of this study is that there are two modes of video honesty check. This makes video honesty check detect fake media better.

The study results show that video authenticity check can be used to develop a stronger video for a deep fake detector. This makes video honesty check a good foundation for developing a better video for a deep fake detector in the future. However, there are still several things to research. First, the system should be tested with different videos to improve video honesty check and to find out how to trick the system to check for fake media. Another thing to research is to develop a video detector. It would be good to test the system with different language sounds. This is because lip movement varies depending on sound.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3420-3427
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0442

References

[1]. Chung, J. S., & Zisserman, A. (2016, November). Out of time: automated lip sync in the wild. In *Asian conference on computer vision* (pp. 251-263). Cham: Springer International Publishing.

[2]. Hariprasad, Y., Iyengar, S. S., & Subramanian, N. (2024). Deepfake video detection using lip region analysis with advanced artificial intelligence based anomaly detection technique. Authorea Preprints.

[3]. Li, Y., Chang, M. C., & Lyu, S. (2018, December). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International workshop on information forensics and security (WIFS) (pp. 1-7). Ieee.

[4]. Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE.

[5]. Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020, December). Detecting deep-fake videos from appearance and behavior. In 2020 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

[6]. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[7]. Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

[8]. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).

[9]. Datta, S. K., Jia, S., & Lyu, S. (2025). Detecting lip-syncing deepfakes: Vision temporal transformer for analyzing mouth inconsistencies. arXiv preprint arXiv:2504.01470.

[10]. Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). Deepfake audio detection via MFCC features using machine learning. IEEE Access, 10, 134018-134028.

[11]. Zhao, H., Zhou, W., Chen, D., Zhang, W., & Yu, N. (2022). Self-supervised transformer for deepfake detection. arXiv preprint arXiv:2203.01265.