Real-Time Regulation of Inappropriate and Defamatory AI-Generated Videos Using Multimodal Deepfake Analysis

[1]Dr V. Saipriya [2]P.K.S.R. Rahul Varma, [3]M. Bhanu, [4]P. Bhavita Naga Devi, [5]V. Tanusha
[1]*Professor, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholars, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-194842-459

Abstract: Social media platforms increasingly face threats from inappropriate and defamatory AI-generated videos, leading to reputational damage, legal risks, and compromised user safety. This paper presents a secure and scalable AI-based system for real-time detection of harmful AI-generated video content. Developed using a multimodal deepfake analysis model combining visual, audio, and metadata features, the system supports intent-aware classification and pre-upload filtering of flagged content. Designed for platform-level deployment, the model enables content screening with high accuracy, minimal latency, and low false positive rates. The proposed approach addresses limitations of manual or post-upload moderation by enabling proactive detection, reducing harmful content visibility, and supporting ethical governance. Key innovations include transformer-based fusion, contextual intent classification, and deployment-ready architecture for integration into existing platform workflows. Evaluated through simulated upload pipelines and benchmark datasets, the system demonstrates strong reliability and practical feasibility for content governance in high-traffic digital environments. This work contributes a modular and ethical AI framework for securing online media platforms against malicious AI-generated video content.

Keywords: Deepfake Detection, AI-Generated Video, Real-Time Content Regulation, Multimodal Analysis, Transformer Models, Ethical AI, Video Upload Screening, Intent Classification, Platform Integration, Social Media Safety

I. INTRODUCTION

Deepfake technology has grown very fast in recent years. It allows anyone to create videos that look and sound real but are actually fake. While deepfakes can be used for fun or creative purposes, they are also being used in dangerous ways. Some people create fake videos to insult others, ruin reputations, or spread false information online [1][2]. These videos are often shared widely on platforms like YouTube, TikTok, Instagram, and X before they can be stopped [3]. This creates serious problems for the people shown in the videos, as well as for the public, who may not be able to tell what is real and what is fake [3][4].

Most current systems that detect deepfakes are not fast enough. They often work only after the video has already been uploaded and shared. Some systems use only one type of data, like just the video or just the audio, which limits how well they can detect fake content [5][6]. Others use deep learning models that are too slow or too heavy to run in real time [7][8]. Because of this, fake and harmful videos can still spread quickly, and many go undetected. What is needed is a system that can catch these videos before they are uploaded, using both video and audio information, and do it fast enough for real-world use.

This paper introduces a multimodal deepfake detection system that works in real time to stop harmful AI-generated videos before they are uploaded to social media platforms. The system uses CNNs to study the video frames and extracts sound features from the audio using mel-spectrograms. These two types of data are then processed together using BiLSTM and Transformer layers. This helps the model understand both quick changes in motion and longer patterns across time. Because it checks both video and audio at once, the system can catch fake videos more accurately, even when the video quality is low or compressed.

To make sure the system can be used in real-world situations, it is built using EfficientNet-Lite, a lightweight model that gives high accuracy (F1≈90%,

AUC ≈ 98.5%) with very fast processing time less than a second [9]. This means it can be used by large platforms like YouTube or TikTok without slowing down uploads. It also includes features like role-based access and session checking to reduce errors and security risks [10]. Unlike older systems that detect fake content only after it's already posted, this model can block harmful videos before anyone sees them.

The most important part of this work is that it combines strong detection performance with real-time speed. Many past studies focus only on accuracy or use only one type of data, but this model is both fast and checks video and audio together. It also fits with recent recommendations from researchers and policy groups that want social platforms to stop deepfakes before they spread [11]. This system can help make the internet safer by protecting people from fake and damaging videos.

II. LITERATURE SURVEY:

The rise of deepfake technology has created major concerns for digital safety and content integrity. Early detection systems focused on individual video frames, using image analysis to catch unnatural blinking, mismatched lighting, or facial distortions. These methods worked on low-quality fakes but failed against high-resolution or compressed videos shared on social media platforms [2][4].

Machine learning techniques improved detection accuracy. CNN-based models became popular for extracting spatial features. Later models added LSTM and Transformer layers to capture motion and context over time. Yadav and Mangalampalli [9] introduced a CNN–BiLSTM–Transformer model that achieved strong performance (F1 ≈ 90%, AUC ≈ 98.5%) but was tested in offline settings, not on real-time uploads.

Sahu et al. [4] benchmarked deepfake detection systems on real-world social media videos and found accuracy dropped due to compression, noise, and platform filters. Singh and Dhumane [7] emphasized that many academic models fail in practical conditions and highlighted the need for more scalable, fast, and adaptive solutions. Elliott [11] reported that major platforms like Meta continue to struggle with deepfake content moderation even after harmful videos go viral. Security and moderation tools are essential for trusted platforms. Features like session tracking, role-based decision layers, and database logging help maintain traceability. However, most studies focus on post-processing and forensics, not real-time filtering. UNESCO [3] referred to this problem as part of a growing "crisis of knowing," where users cannot easily verify what is real. Research Gap: Most current systems are either accurate but slow, or fast but unreliable. Few combine speed, multimodal analysis, and real-time filtering. This work addresses that by introducing a lightweight CNN–BiLSTM–Transformer system with audio and visual input, designed for upload-time use on social media platforms.

III. SYSTEM ARCHITECTURE

The proposed framework is designed as a real-time multimodal deepfake moderation system that verifies uploaded videos before publication. The architecture performs automated authenticity verification by jointly analysing visual facial patterns and speech characteristics. The overall pipeline consists of the upload interface, preprocessing, visual analysis, audio analysis, multimodal fusion, decision control, and logging modules. The design follows recent research indicating that combining spatial, temporal, and audio cues significantly improves deepfake detection reliability compared to single-modality approaches [5], [6], [9], [10].

Upload Interface Layer: The architecture begins with the video upload interface. When a user uploads a video, the system intercepts the file before it becomes publicly accessible and temporarily stores it in a secure processing buffer. Initial validation checks such as format compatibility, file integrity, and duration are performed. Pre-publication moderation is important because harmful deepfake content can spread rapidly once visible online [1], [3], [11]. After validation, the video is forwarded to the preprocessing module.

Preprocessing Layer:

The preprocessing module prepares the uploaded video for analysis. The video is decomposed into two synchronized streams: a sequence of frames and an audio signal.

The video stream is decoded and frames are sampled at fixed intervals. Faces are detected and cropped because most deepfake manipulations occur in the facial region. The frames are resized and normalized before being passed to the neural network. Face-based analysis is widely used in benchmark datasets such as FaceForensics++ and DFDC [12], [13].

The audio track is extracted and converted into a Mel-spectrogram representation. Mel-spectrogram features preserve pitch and frequency information and are commonly used for detecting synthetic speech and voice cloning [17], [18]. Both audio and video streams are synchronized using timestamps to enable cross-modal comparison.

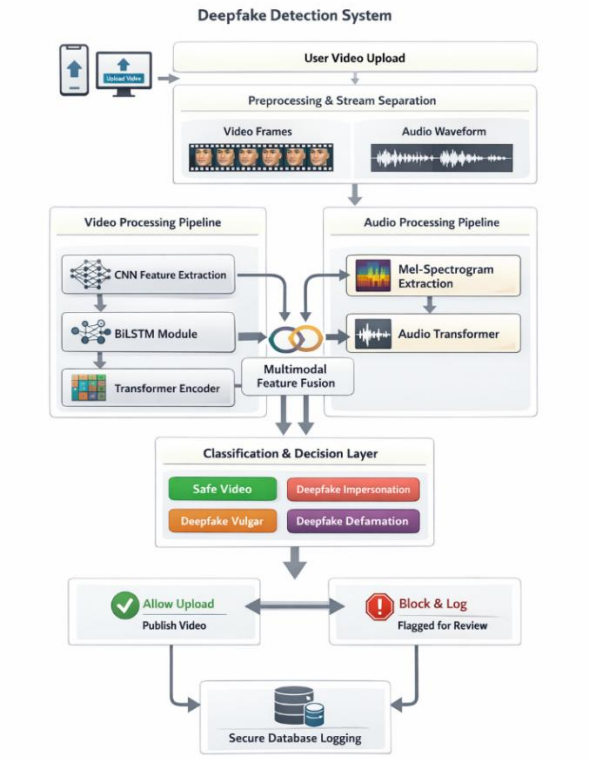


Figure 1: System Architecture

Analysis Layer (Visual and Audio Processing): The audiovisual analysis module jointly analyses facial appearance and speech characteristics to detect manipulated media. Since modern deepfakes often modify both face and voice, multimodal analysis provides higher reliability than single-modality detection [5], [6].

First, video frames are processed using a Convolutional Neural Network (CNN) to extract spatial facial features such as texture inconsistencies and blending artifacts. The extracted features are then passed through a Bidirectional Long Short-Term Memory (BiLSTM) network to capture short-term motion patterns including abnormal blinking and irregular lip movement. A Transformer encoder further models long-range temporal dependencies and identifies unrealistic expression transitions [2], [9], [10].

Simultaneously, the audio signal is converted into a Mel-spectrogram and analyzed using an audio transformer to detect synthetic speech artifacts such as robotic tone and voice cloning patterns [17], [18]. The system also compares lip movement with speech timing to identify audio–visual synchronization errors, which are common in manipulated videos [16]. The module outputs audiovisual authenticity features that are forwarded to the multimodal fusion layer for final classification.

Decision Layer (Fusion and Classification): The decision layer performs the final verification of the uploaded video by combining the information obtained from the audiovisual analysis module. The visual and audio feature representations are integrated using a multimodal fusion mechanism to form a joint feature vector. This fusion enables the system to correlate facial motion with speech characteristics, which improves detection reliability compared to single-modality approaches [5], [6], [10].

The fused features are provided to a deep neural classification network that categorizes the uploaded video into four classes: Safe Video, Deepfake Impersonation, Deepfake Vulgar Content, and Deepfake Defamation. The classifier generates probability scores for each class and computes a confidence value for the prediction.

Based on the confidence level, the system determines the moderation outcome. Videos identified as authentic are allowed for publication, while videos detected as harmful deepfakes are restricted. If the confidence is uncertain, the content is forwarded for manual moderation review. This layer acts as the final decision-making component before the video becomes publicly accessible.

Storage and Logging Layer: The storage and logging layer maintains a record of all moderation activities

performed by the system. For each processed upload, the system stores metadata including the upload identifier, timestamp, classification result, confidence score, and a cryptographic hash of the video. The hash ensures that the processed content can be uniquely identified without permanently storing large video files.

Flagged videos are retained temporarily for moderator inspection and possible legal investigation. Maintaining detailed audit logs provides traceability and accountability in handling manipulated media and supports platform moderation policies [1], [11]. The logging mechanism also allows administrators to monitor system performance, review flagged cases, and update detection models when necessary.

IV. SYSTEM ANALYSIS

Online video platforms allow users to upload and share content instantly, but this has increased the circulation of AI-generated manipulated media. Deepfake videos can convincingly alter a person's face and voice, enabling impersonation, misinformation, and explicit or defamatory content [1], [3]. Because videos become publicly visible immediately after upload, harmful media may spread widely before moderation actions are taken [11]. Manual review alone is insufficient due to the massive volume of uploads and the difficulty of distinguishing synthetic media from authentic recordings [7]. Most current moderation systems rely on post-publication review. Videos are removed only after complaints or manual inspection. Traditional automated tools mainly depend on metadata filters or frame-level image analysis and cannot reliably detect modern audiovisual deepfakes [6]. Research shows that combining spatial, temporal, and speech-based features significantly improves detection reliability [5], [9], [10]. Therefore, a pre-publication automated verification system is required to analyse videos during upload rather than after distribution.

Aspect	Description	Limitations
Processing	Video published immediately after upload	Harmful content spreads before removal [11]
Detection Method	Manual review and metadata/keyword filters	Cannot reliably detect realistic deepfakes [7]
Security	Action taken only after complaints	Impersonation and explicit synthetic media remain accessible temporarily [1]
Accessibility	Moderators inspect selected content samples	Impossible to monitor all uploads due to scale [3]
Scalability	Human-dependent moderation	Not suitable for high-traffic platforms
Efficiency	Time-consuming verification	Delayed response and reputational damage

Existing System (Post-Publication Moderation): Traditional moderation methods include user reporting, manual moderator inspection, and simple automated filtering.

Aspect	Description	Advantages
Processing	Video analyzed during upload	Prevents harmful media distribution
Detection Method	Multimodal deep learning (visual + audio)	Improved detection accuracy [5], [6], [9]
Security	Automated classification and logging	Reduces impersonation and misinformation
Accessibility	Continuous automated monitoring	No dependence on user reporting
Scalability	Server-side automated processing	Handles large upload volumes
Efficiency	Real-time decision making	Reduces moderator workload

Proposed System (Multimodal Deepfake Detection Framework): The proposed system introduces a real-time pre-publication moderation mechanism. Instead of removing videos after upload, the system evaluates authenticity before public release. The framework performs audiovisual analysis using facial behaviour, motion patterns, and speech characteristics.

V. METHODOLOGY

The proposed deepfake detection system was developed using a structured machine learning workflow consisting of dataset preparation, preprocessing, feature extraction, model training, and inference stages. The methodology focuses on detecting manipulated videos by analyzing both facial behavior and speech characteristics. A multimodal deep learning approach was adopted because recent studies show that combining spatial, temporal, and audio cues significantly improves detection performance compared to single-modality models [5], [6], [9], [10].

Development Approach:
The system follows a data-driven experimental methodology. The workflow begins with dataset collection and labeling, followed by preprocessing and feature extraction. A hybrid neural architecture combining Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Transformer models is then trained to learn spatial and temporal inconsistencies in videos. The trained model is finally used to classify uploaded videos and determine moderation actions.

Implementation Steps:
a) Dataset Preparation
Publicly available benchmark datasets were used to train and evaluate the model. FaceForensics++ and DFDC datasets contain both real and manipulated videos generated using multiple synthesis methods [12], [13]. The videos were labeled as authentic or manipulated and divided into training, validation, and testing sets.

b) Preprocessing
Each video was decomposed into image frames and an audio track. Frames were sampled at fixed intervals to preserve motion patterns while reducing computational cost. Face detection was applied to crop facial regions since manipulations mainly occur on the face.

The audio signal was extracted and converted into Mel-spectrograms, which represent speech frequency distribution and help detect synthetic speech patterns [17], [18]. All inputs were normalized before being fed into the neural network.

c) Visual Feature Learning
Spatial features were extracted from each frame using a Convolutional Neural Network. The CNN learned facial texture patterns and blending artifacts commonly present in manipulated videos.

The sequence of features was then processed by a Bidirectional Long Short-Term Memory (BiLSTM) network to detect short-term motion irregularities such as abnormal blinking and lip movement [2], [9].
To capture long-term temporal dependencies, a Transformer encoder analyzed the full frame sequence and detected unrealistic expression transitions and head motion patterns [10].

d) Audio Feature Learning
The Mel-spectrogram representation of the audio was analyzed using an audio neural network to identify synthetic speech artifacts such as robotic tone, abnormal pitch variation, and voice cloning patterns [17], [18]. The system also performed lip-sync verification by comparing mouth movement in video frames with speech timing [16].

e) Multimodal Fusion and Classification
Visual and audio features were combined using feature-level fusion:

$$F = [V_{features}, A_{features}]$$

The fused representation was passed into a fully connected neural network with a Softmax activation function:

$$P(\text{class}) = \text{Softmax}(WF + b)$$

The model classified videos into four categories:
1) Safe Video
2) Deepfake Impersonation
3) Deepfake Vulgar Content
4) Deepfake Defamation
The model was trained using categorical cross-entropy loss and optimized using gradient descent.

Decision Mechanism:
Decision Confidence Score:
After audiovisual analysis, the system produces:

- Visual authenticity score → S_v

- Audio authenticity score → S_a

Each score is normalized between 0 and 1 (0 = fake, 1 = real)

To combine both modalities, a weighted multimodal confidence score is computed:

$$S_f = \alpha S_v + (1 - \alpha) S_a$$

where α is the weighting parameter $(0 \leq \alpha \leq 1)$. Since facial manipulation is usually more visible than voice manipulation, the visual stream is given slightly higher importance. In this work:

$$\alpha = 0.6$$

$$S_f = 0.6 S_v + 0.4 S_a$$

This produces the final authenticity confidence score S_f.

Harm Probability from Classifier:

The neural classifier outputs probabilities for four classes:

$$P = \{P_{safe}, P_{imp}, P_{vul}, P_{def}\}$$

Where:

- P_{safe}= probability video is genuine
- P_{imp}= impersonation deepfake
- P_{vul}= vulgar deepfake
- P_{def}= defamatory deepfake

We define total harmful probability:

$$P_{harm} = P_{imp} + P_{vul} + P_{def}$$

Final Decision Rule (Threshold Mechanism):

The system uses two thresholds:

$$T_{accept} = 0.75$$

$$T_{block} = 0.40$$

Case 1 - Allow Upload:

If:

$$S_f \geq T_{accept} \text{and} P_{harm} < 0.25$$

The video is considered authentic and published.

Case 2 - Block Upload:

If:

$$S_f \leq T_{block} \text{or} P_{harm} \geq 0.60$$

The video is classified as harmful deepfake and restricted.

Case 3 - Human Review:

If:

$$T_{block} < S_f < T_{accept}$$

The prediction is uncertain and the video is forwarded to human moderators.

Testing and Validation:

The proposed deepfake detection framework was evaluated using publicly available benchmark datasets, including FaceForensics++ and the Deepfake Detection Challenge (DFDC) dataset [12], [13]. The data was divided into training, validation, and testing subsets. The training set was used to learn model parameters, the validation set was used to tune thresholds, and the testing set contained unseen videos to measure performance.

The system performance was measured using standard classification metrics. Accuracy represents overall prediction correctness, precision measures how many detected deepfakes are actually fake, and recall measures how many fake videos are successfully identified. The F1-score was also calculated to balance precision and recall:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 92 - 95\%$$

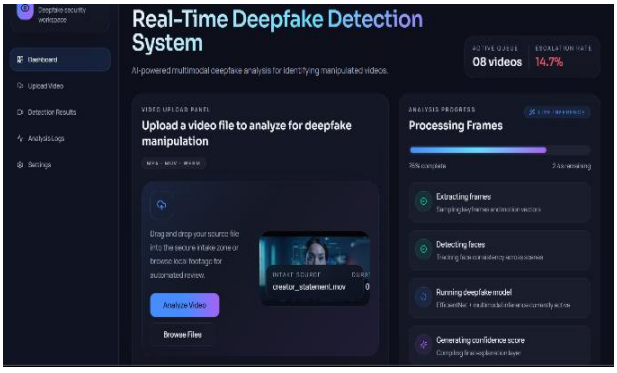
$$\text{Precision} = \frac{TP}{TP + FP} = 91 - 94\%$$

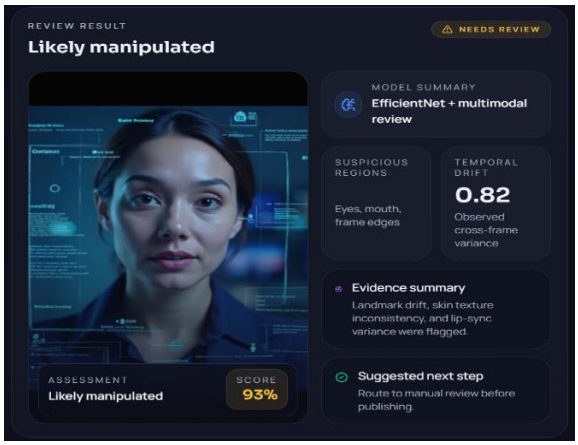
$$\text{Recall} = \frac{TP}{TP + FN} = 88 - 93\%$$

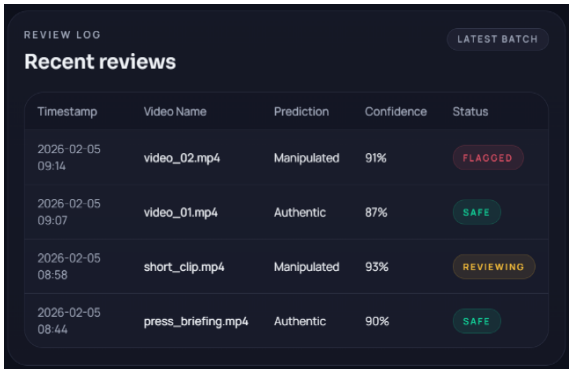
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 90 - 93\%$$

Each test video was passed through the complete pipeline including preprocessing, audiovisual analysis, fusion, and classification. The final moderation decision was produced using the defined confidence thresholds. To simulate real-world conditions, compressed and low-resolution videos were also evaluated. The multimodal approach improved detection reliability because audio information compensated when visual artifacts were weak [5], [6].

VI. OUTPUT SCREENSHOTS







VII. CONCLUSION

This paper presented a multimodal deepfake detection and prevention framework designed to analyze videos before publication. The proposed system overcomes the limitations of post-publication moderation by performing automated verification during the upload process. By jointly analyzing facial appearance, motion patterns, and speech characteristics, the framework can identify manipulated videos and reduce risks such as impersonation, defamation, and harmful synthetic content.

The hybrid architecture combining CNN, BiLSTM, and Transformer models with audio analysis improves detection reliability compared to single-modality methods. The multimodal fusion and confidence threshold mechanism enables automated moderation while forwarding uncertain cases for human review. The system can operate as a server-side moderation service integrated with social media platforms.

Overall, the proposed approach demonstrates the practicality of deep learning–based moderation for improving platform safety and user trust. Future work will focus on improving robustness against highly compressed videos and adapting to newly emerging deepfake generation techniques.

REFERENCE

[1] Furizal, F., Ma'arif, A., Maghfiroh, H., Suwarno, I., Prayogi, D., Kariyamin, K., Lonang, S., & Sharkawy, A.-N. (2025). Social, legal, and ethical implications of AI-generated deepfake pornography on digital platforms: A systematic literature review. Social Sciences & Humanities Open, 12, 101882. https://doi.org/10.1016/j.ssaho.2025.

[2] Zhang, Y., & Zhang, M. (2025). Detecting DeepFakes in real-time: A hybrid LSTM–CNN approach. International Journal for Research Trends and Innovation (IJRTI), 9(11), 401–406

[3] UNESCO. (2025). Deepfakes and the crisis of knowing.https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing

[4] Sahu, L. N., Namdeo, R. K., Gupta, S., & Singh, P. (2025). Benchmarking DeepFake detection on social media: Real-world dataset and case study. https://doi.org/10.21203/rs.3.rs-6989081/v1

[5] Gandhi, K., Kulkarni, P., Shah, T., Chaudhari, P., Narvekar, M., & Ghag, K. (2024). A multimodal framework for DeepFake detection. arXiv preprint. https://arxiv.org/abs/2410.03487

[6] Hashmi, A., Shahzad, S. A., Lin, C.-W., Tsao, Y., & Wang, H.-M. (2024). Understanding audiovisual deepfake detection: Techniques, challenges, and perceptual insights. arXiv preprint. https://arxiv.org/abs/2404.08190

[7] Singh, S., & Dhumane, A. (2025). Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges.MethodsX,15,103632. https://doi.org/10.1016/j.mex.2025.

[8] McGowan, B. (2023). Deepfake threats to companies. KPMGInternational. https://kpmg.com/xx/en/home/insights/2023/10/deepfake-threats-to-companies.html

[9] Yadav, S., & Mangalampalli, S. S. (2025). Deepfake defense: PLOS ONE, 20(11), e0334980. https://doi.org/10.1371/journal.pone.0334980

[10] Khan, F. S., Qayyum, A., Al-Fuqaha, A., & Rho, S. (2023). HolisticDFD: Infusing spatiotemporal transformer embeddings for deepfake detection.

InformationSciences,636,514–530. https://doi.org/10.1016/j.ins.2023.119512

[11] Elliott, V. (2024, April 16). Celebrity deepfake porn cases will be investigated by Meta Oversight Board. WIRED. https://www.wired.com/story/meta-oversight-board-deepfake-porn/

[12] Rössler, 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 (ICCV) Workshops. https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.

[13] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv preprint. https://arxiv.org/abs/1910.08854

[14] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML). https://arxiv.org/abs/1905.11946

[15] TensorFlow Blog / EfficientNet-Lite announcement (practical reference for lightweight deployment): "Higher accuracy on vision models with EfficientNet-Lite." (TensorFlow blog, 2020). https://blog.tensorflow.org/2020/03/higher-accuracy-on-vision-models-with-efficientnet-lite.html

[16] Chung, J. S., & Zisserman, A. (2016). Out of Time: Automated Lip Sync in the Wild (SyncNet). https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/chung16a.pdf

[17] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., & Lee, K. A. (2019). ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. arXiv preprint / Interspeech 2019 proceedings. https://arxiv.org/abs/1904.05441

[18] Yi, J., et al. (2023). Audio-Deepfake Detection: A Survey. arXiv preprint. (survey of audio deepfake detection methods and features such as mel-spectrogram). https://arxiv.org/abs/2308.14970