International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026

Multimedia Forgery Detection Using CNN: Identifying Fake Visuals and Dubbed Audio in Videos

Reshma T R
Department of Artificial Intelligence
Jain (deemed-to-be) University,
Bangalore, India
jupg25mtech21484@jainuniversity.ac.in

Janaki Kandasamy
Department of Computer Science and Engineering
Jain (deemed-to-be) University,
Bangalore, India
k.janaki@jainuniversity.ac.in

Abstract - *Synthetic media enabled by generative AI has become increasingly realistic making it highly challenging to differentiate altered images, video, and audio. The majority of the existing detection algorithms only classify alterations in a single modality at a time, limiting the system to identify partially altered content, such as a real image with fake audio or a fake video with real audio. The paper proposes "DeepGuard" a comprehensive multimedia fraud detection system for images, videos and audio. The system consists of a cross-modal consistency module to evaluate the consistency of visual and audio segments of a video. The system uses EfficientNet-B0 to analyze both images and videos, and it uses a four-layer CNN on Mel spectrograms to classify audio. Analyses of the video results are integrated with the audio analysis for each video and are processed using a lip-synchronization test to assess the category of manipulation that occurred. Experimental results on benchmark datasets achieved validation accuracies of 98.17% for image detection, 98.40% for video detection, and 99.86% for audio detection. The system also generates a timestamped forensic report summarizing the forgeries detected and suggested actions.*

Keywords- deepfake detection, multimedia forgery, CNN, cross-modal analysis, lip-sync detection, Grad-CAM, Mel spectrogram, audio dubbing, digital forensics

I. INTRODUCTION

Advances in generative media tools have shifted from requiring professional software and trained professionals to only requiring downloading consumer software with no technical expertise to perform face swapping, synthesize voice, or generate a realistic synthetic portrait. These technologies can be misused to produce fabricated videos of a speech, synthetic audio recordings authorizing financial transactions, and manipulated video footage presented as legal evidence. A survey documented instances of synthetic media-based fraud from 2022 to 2024, indicating approximately a fourfold increase in such attempts [1].

The study of forgery detection is equally fragmented across multiple modalities as the technologies used to generate them. Visual detection systems are trained to detect artifacts and irregularities associated with face-swapped subjects in image and video frames [2][3]. Audio detection systems are developed using synthetic speech training data to determine differences between text-to-speech distinct from natural speech recordings by evaluating their corresponding frequency domain features [4].

Existing studies have developed effective detection systems for different modalities, yet cross-modal attacks are not adequately addressed. If the facial content in the video is authentic, but the voice is synthetic, and the visual-only detector is applied, the video may be incorrectly reported as authentic. An audio detector running on the extracted audio would detect the synthetic audio. However, it only functions when explicitly performed. The same asymmetry applies in the other direction, i.e., when the voice is authentic but facially swapped visual content and the audio is evaluated with an audio detector, which may identify no irregularity, whereas a visual detector would identify manipulation.

Another limitation of existing studies is the limited explanatory details. Most systems produce confidence scores with real or fake classification. While it is suitable for large-scale filtering it remains inadequate during forensic analysis, where the analyst must determine how and why evidence was manipulated and identify supporting indicators supporting the conclusion.

Both of the previously identified limitations are addressed in this study. DeepGuard is a web application for assessing images, videos and audio files using a trained deep learning model. It conducts a cross-modal consistency analysis on video inputs by comparing audio and visual predictions and assessing the degree to which lip movement matches audio energy patterns. The integrated analysis of all three inputs produces a classification of whether it is a full deepfake, dubbed audio, swapped face or authentic. In addition, each analysis generates a timestamped forensic report. The major contributions of this study can be defined as:

i. An integrated system for identifying images, video and audio within a single browser-based platform.
ii. A cross-modal consistency analysis for visual audio and lip-synchronize data to detect partial manipulation, including audio and face-swapped videos.
iii. The application of Grad-CAM visualization of image predictions of specific facial regions contributing to the classification decision.
iv. Automated generation of forensic reports for every analysis session, structured in a human-readable format.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026

II. Related Work

A. Existing Systems

1) Detecting Manipulated Images and Video Frames

The FaceForensics++ benchmark [2] has been the benchmark for visual fraud detection since its launch in 2019, demonstrating that through training binary classification models using frame-level characteristics, a high level of accuracy can be achieved in detecting manipulated content in the dataset, which consists of video clips affected by four distinct manipulation techniques. Furthermore, Rössler et al. showed that the XceptionNet architecture performs exceptionally well. Li et al. introduced a new dataset called Celeb-DF [3], which was designed to include a larger quantity of high-resolution manipulated samples of celebrities and a more challenging benchmark than the previous datasets. The results from Celeb-DF were consistent with the earlier studies as the models trained on FaceForensics++ showed reduced accuracy when evaluated on unseen manipulation methods. This limitation persisted even after employing approaches independent of the specific generation techniques, such as frequency domain characteristics and attention. However, none of these algorithms determines whether an arbitrary audio stream is real, as they solely generate predictions from visual data.

2) Detecting Synthetic Speech

Since 2015, research in the field of anti-spoofing for automatic speaker verification (ASV) has been guided through the ASVspoof evaluation initiative [4]. For the 2019 benchmark cycle, seventeen distinct text-to-speech (TTS) synthesis and voice conversion (VC) systems were employed to generate training and evaluation datasets. These datasets were utilized in the present study. The challenge established several competitive baseline models and adopted the Equal Error Rate (EER) as the benchmark for evaluation. The AASIST model [5], which considers the spectrogram as a graph and employs a spectro-temporal attention mechanism, achieved the highest error rate for the ASVspoof 2019 logical access section among all published systems. In parallel, Müller et al. [6] assessed whether anti-spoofing systems developed on ASVspoof data could generalize to naturally recorded audio, and discovered a substantial performance gap, indicating that benchmark conditions do not adequately represent the distribution of synthetic audio in a real-world environment.

3) Joint Audio-Visual Analysis

FakeAVCeleb [7] was the first benchmark dataset to provide a paired face-swapped video and cloned audio samples for training multimodal detectors and released in 2021. AVoiD-DF [8] proposed a collaborative learning paradigm aligning the audio and visual modalities within a shared representation space. The LAV-DF dataset [9] shifted the task objective from binary classification to temporal localization by requiring models to identify the temporal segments within a clip where manipulation occurred instead of assigning a label to the complete recording. Deepfake-Eval-2024 [1] evaluated multiple existing detectors by testing them on data gathered from social media in 2024. When compared to benchmark performance metrics from the original training data the performance of these detectors declined significantly on the newly collected real-world data. There are no automated forensic reports produced by any of these systems, manipulation type classification based on both dubbing and face-swapping, or an end-user interaction platform.

B. Proposed System

DeepGuard has a cross-modal consistency component that none of the previous systems included, but this system integrates three separate detection processes into a unified interface. The image pipeline includes a Grad-CAM-based interpretability module that explains the predictions of a pre- trained and fine-tuned convolutional neural network classifier trained on balanced datasets of real and fake face images. Finally, the video pipeline processes the outputs from the sampled video frames and aggregates them through a majority voting process to produce a single output for each video frame. After the audio is transformed to a Mel spectrogram, the audio pipeline processes the recording through a custom convolutional neural network (CNN) trained on the ASVspoof 2019 dataset. The three audio and video processes are executed simultaneously for cross-modal analysis. The outputs are forwarded to a consistency module to compute lip-synchronization score and classify the inputs into one of five manipulation categories based on visual prediction, audio prediction and synchronization score. The overall architecture is shown in fig. 1.

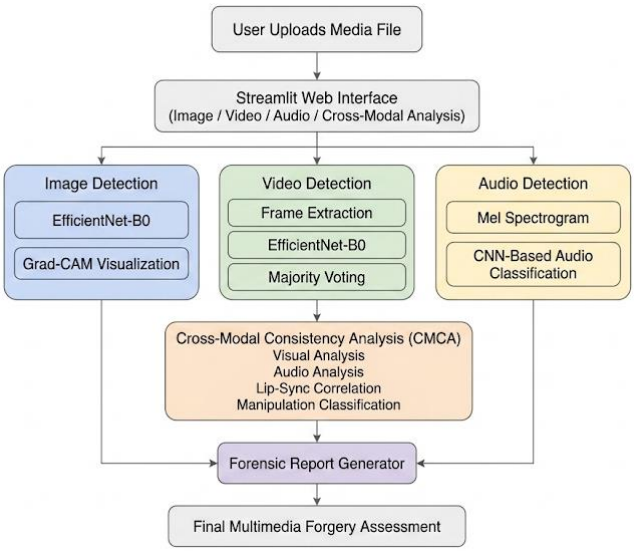


Fig. 1. DeepGuard System Architecture

TABLE I
DeepGuard System Component Summary

Module	Model	Dataset	Accuracy
Image	EfficientNet-B0 + Grad-CAM	Mendeley 2025	98.17%

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026

Video	EfficientNet-B0 (frames)	SDFVD 2.0	98.40%
Audio	Custom 4-block CNN	ASVspoof 2019	99.86%
CMCA	Signal processing	SDFVD 2.0	Rule-based

III. METHODOLOGY

A. Application Structure

The application operates through a web browser implemented using Streamlit as the application framework. The application has four operational modes: image detection, video detection, audio detection, and cross-modal analysis accessible to the user can access. When the user selects a mode, the corresponding interface panel is displayed enabling the user to upload a file for analysis. Each analysis result presents a forensic report of the analysis with a unique session identifier, a timestamp, the manipulation type detected and a recommendation follow up action. Supporting evidence is also provided in the form of a heatmap visualization for images, a grid of annotated frames for videos, a waveform and a spectrogram visualization for audio files.

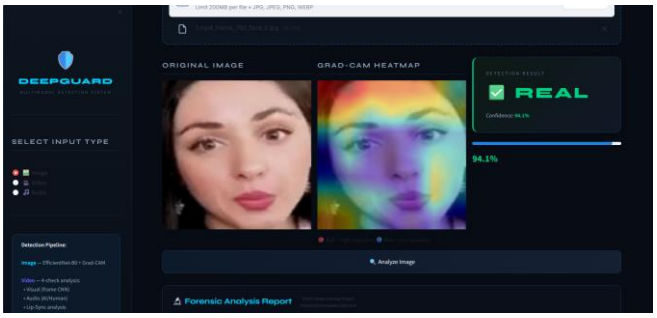
The project is composed of a main application file with three model files and three utility files. The model files contain architecture definitions and functions required for inference operations. The utility files handle lip synchronization measurement, cross-modal scoring, and report generation. This modular design simplifies the replacement of the model files with a retrained version through a consistent file interface without modifying the interface code within the application file.

B. Image Detection with Grad-CAM Explainability

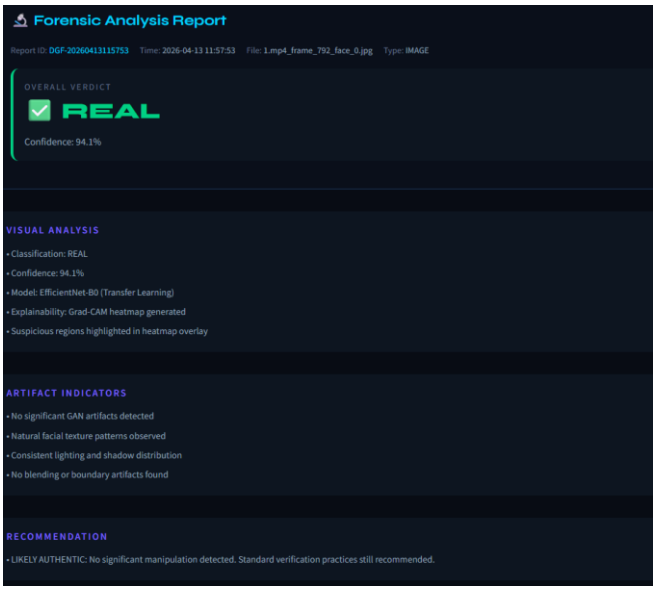
The EfficientNet-B0 model was chosen for both image and video pipelines as it offers a balance between accuracy and the amount of processing power available on CPUs. Initially, the model is trained on the ImageNet classification dataset to classify images. The original architecture was adapted to have a two-layer classifier with batch normalization and higher dropout rates of 0.5 and 0.4 for each additional layer, respectively. As a result, the additional layers introduce stronger regularization to address the skewed class distribution of the training data. While over 100,000 frames of video frames, fewer than 4,000 frames were real content, while the remaining samples consist of manipulated frames. In the absence of an additional dropout regularization, the network rapidly converges to predicting the fake class for all inputs. Balanced sampling addresses imbalanced class issues by selecting an equal number of real and synthetic samples within a batch to ensure balanced learning on both classes.

The AdamW optimization algorithm is employed during training with a learning rate that decreases gradually with a weight decay of 0.01 and a cosine annealing schedule. To prevent overconfident predictions, a cross-entropy loss function with a smoothing factor of 0.1 is applied. If the evaluation accuracy fails to improve over five consecutive epochs, the training process is terminated based on an early stopping criterion.

The Grad-CAM approach is utilized to generate an interpretation for a prediction. The backpropagated gradients are integrated with the feature map activations at the final convolutional layer to generate a visual map highlighting the relative importance of different image regions in the classification decision. Regions with high gradient activity correspond mostly strongly to the predicted class score, with low gradient intensity having minimal influence. Regions of the input image with high gradient intensity appear in warm colour tones for highly influential areas, and regions with low gradient intensity appear in cool colour tones for less significant regions, with the heatmap resized to fit the resolution of the input image. When a face is classified as false, the heat map displays increased activity around the facial boundary, or where blending artifacts occur, such as the hairline region or the areas near the ears as shown in the fig. 2a. and fig. 2c.

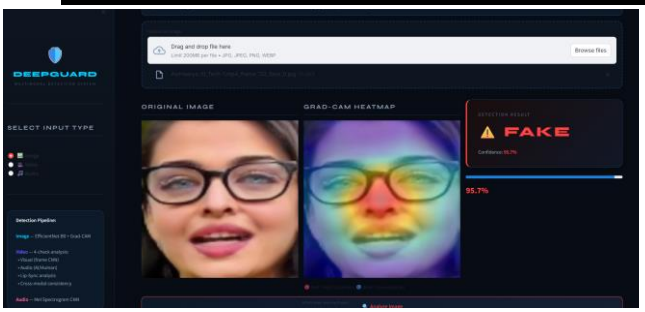


2a. Real image prediction with Grad-CAM

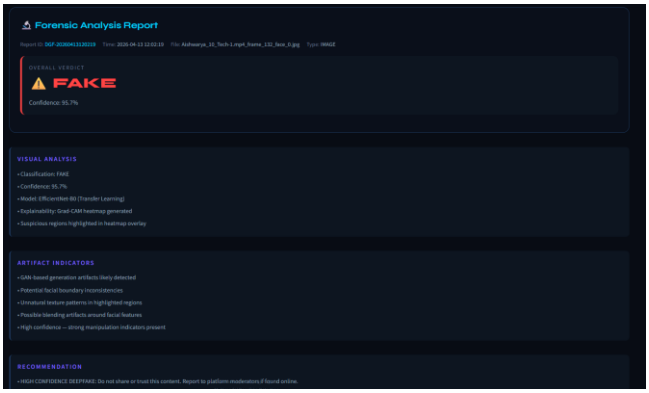


2b. Generated forensic report for authentic image

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026



2c. Fake image prediction with Grad-CAM



2d. Generated forensic report for manipulated image

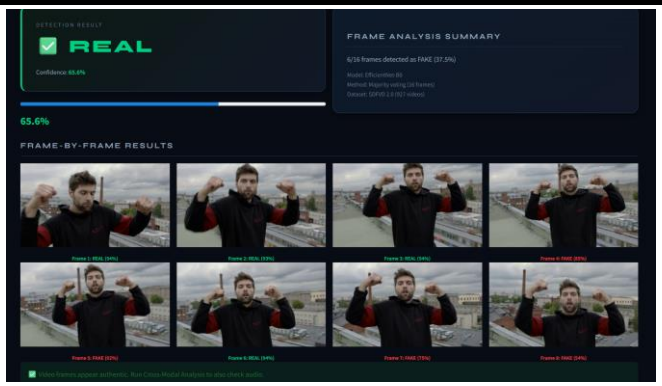
Fig. 2. Image forgery detection results with Grad-CAM and forensic reports.

C. Video Detection Module

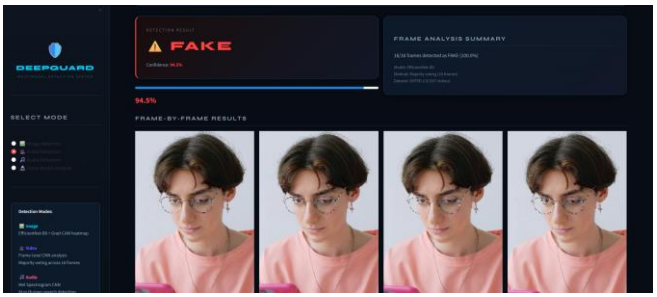
For video analysis, OpenCV extracts sixteen uniformly distributed frames from the uploaded clip. Each frame is independently evaluated with the EfficientNet-B0 architecture used for image detection pipeline. The final prediction is obtained through majority vote. If the majority of sampled frames are classified as fake, the video is classified as manipulated. The confidence value is measured through the mean fake prediction score across all the frames.

During the training phase, ten frames were extracted from each video of the SDFVD 2.0 dataset as showing in fig. 3a and 3b. The effective number of training instances is ten times larger than a single training sample. Given that the SDFVD 2.0 dataset contains only 927 videos, the additional samples significantly improve the performance of the model. To prevent the model from positional bias, extraction locations are randomized during training. Additionally, the extracted frames are processed through the same augmentation methods as the image dataset, including colour jitter, rotation, and flipping.

The model is fine-tuned by updating the last two blocks of the network while all the preceding layers remain frozen. These earlier ImageNet pretrained layers serve as a source of generic visual representations. Restricting the updates to a limited number of trainable parameters reduces the risk of overfitting. This approach is referred to as partial fine-tuning.



3a. Real video detection result



3b. Fake video detection result

Fig. 3. Video forgery detection results for authentic and manipulated videos.

D. Audio Detection Module

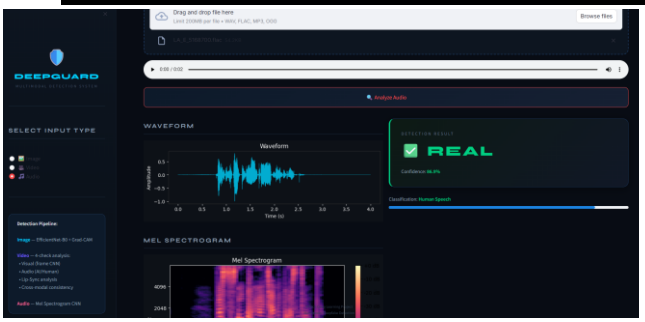
For audio processing, the raw waveform is converted into a Mel spectrogram instead as the primary input representation as shown in fig. 4a. and fig. 4c. A Mel spectrogram captures the temporal intensity of an audio sample across frequency bands. The audio is resampled to 16kHz and processed using a 1024-point short-time Fourier transform with a hop size of 512 samples. Finally, the resulting spectrogram has 128 frequency bins and is treated as a single-channel image for the CNN-based classifier.

The classifier comprises four blocks, each consisting of a convolution, batch normalization, ReLU activation, a max-pooling layer, and two-dimensional dropout. Mild regularization is applied in the earlier layers, where the learned features are more generic, while stronger regularization is applied in the deeper layers. Consequently, the dropout rates increase in deeper layers. The fully connected classifier employs dropout rates of 0.5 and 0.4. Since input data is a single-channel spectrogram image, which differs from RGB image inputs used in pretrained vision models, the neural network was designed from scratch.

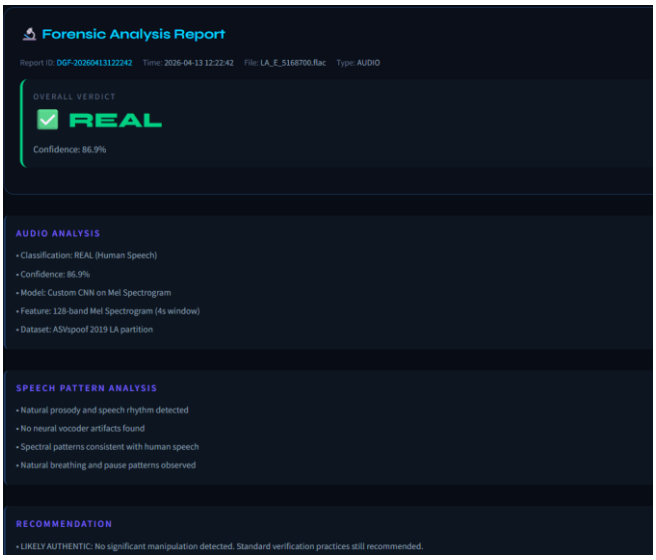
In the ASVspoof 2019, the logical access partition contains an imbalance of class distribution, with the fake samples outnumbering real samples by roughly 9:1. The entire dataset is balanced equally across both classes during training. In addition, the validation set is similarly balanced to ensure that early stopping decisions and the accuracy score are not biased toward the majority class.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026

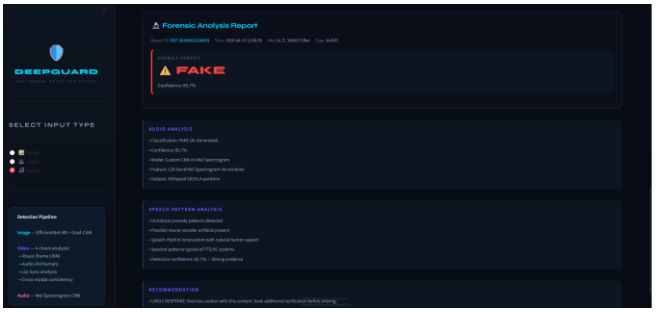


4a. Real audio detection result



4b. Real audio forensic report

4c. Fake audio detection result



4d. Fake audio forensic report

Fig. 4. Audio forgery detection results with spectrogram analysis and forensic reports.

E. Cross-Modal Consistency Analysis (CMCA)

The cross-modal module is invoked when the cross-modal analysis interface is selected for video input. It evaluates three distinct signals (i) a lip-synchronization score computed from visual frames and associated audio waveform (ii) the visual verdict prediction generated by the video detection pipeline. (iii) the audio classification result produced by the extracted audio track using the audio detection pipeline.

1) Lip movement measurement.

Thirty frames are uniformly sampled from the video. For each frame, the OpenCV Haar cascade face detector identifies the face bounding box. The mouth region is extracted from the lower third of this region. The Canny edge detector is applied, and the resulting edge map is used to estimate the edge density through the proportion of pixels classified as edges. The pixel variance of the mouth region is also computed. These two quantities are combined into a single lip openness value per frame. Across the thirty frames, this produces a time series representing how the mouth moved throughout the clip.

2) Audio energy measurement.

To align with the sampled frames, the audio waveform is divided into thirty equally sized windows in order to represent them. The loudness of each segment is computed using the Root Mean Square (RMS) method applied to audio waveform samples within each window. This generates a second time series equal in length to the sampled frame sequence representing the loud audio over time. Before the comparison of the two-time series, both are normalized to the range of 0 to 1.

3) Synchronization score.

The Pearson correlation coefficient is computed between the lip synchrony data and the audio energy sequence. A strong positive correlation indicates that the mouth movements match the loudness of speech and is consistent with the real speech production. A weak or negative correlation suggests the audio and video are not aligned. The coefficient is scaled from 0 to 100. where scores above 55 denote synchronization, below 38 denote mismatch, and intermediate values are labelled uncertain.

4) Manipulation type classification.

The visual, audio, and lip-synchronization outputs are integrated using the decision rule shown in Table II. The audio dubbed category corresponds to real visual content, with synthetic voice, whereas the face swapped category corresponds to real audio paired with manipulated visual content. Both categories capture cross-modal threats that cannot be detected by a single-modality detector independently.

TABLE II

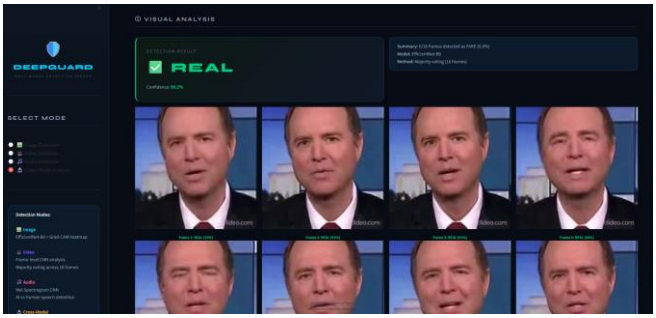
CMCA Manipulation Type Classification

Visual	Audio	Lip-Sync	Verdict
REAL	REAL	SYNCED	AUTHENTIC
FAKE	FAKE	MISMATCH	FULL DEEPFAKE
REAL	FAKE	MISMATCH	AUDIO DUBBED
FAKE	REAL	MISMATCH	FACE SWAPPED

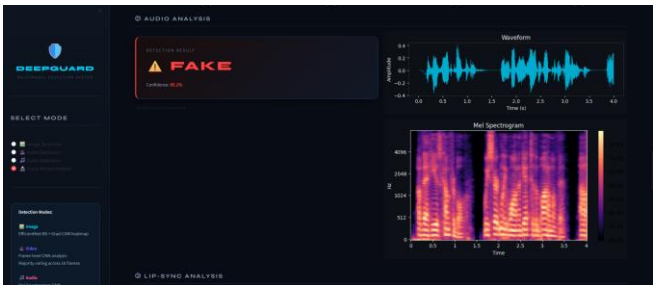
2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026

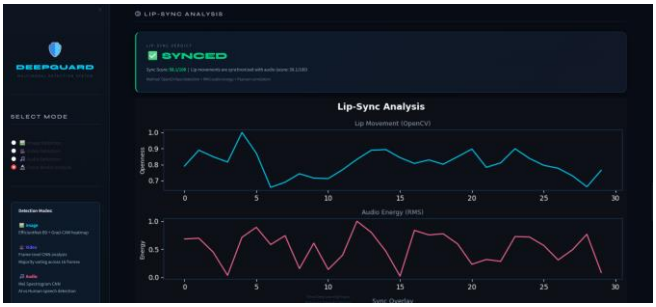
Any	Any	UNCERTAIN	ANOMALY DETECTED



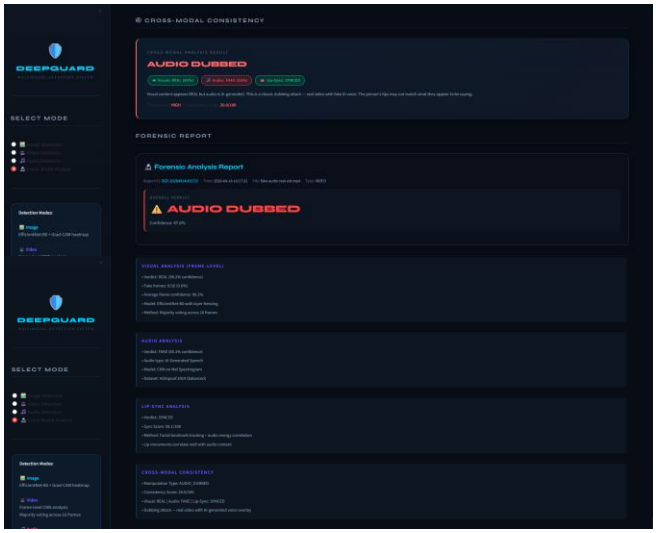
5a. Visual analysis result



5b. Audio analysis result



5c. Lip-sync correlation analysis



5d. Generated forensic report

Fig. 5. Cross-modal analysis outputs

IV. DATASETS

Three publicly available datasets were used across the detection pipelines, each previously cited in peer-reviewed literature. The image pipeline utilized a Mendeley data dataset released in January 2025 [10] comprising 110,694 face frames from 480 videos of 30 subjects. The manipulated frames were generated using Roop Face-Swapper and Akool AI, whereas the real-world recordings were recorded in controlled conditions. Face localization was performed using MTCNN (Multi-Task Cascaded Convolutional Networks), and each face was cropped to 500 x 500 pixels in size. Equal sampling was utilized during training, as the synthetic samples outnumbered the real class samples by approximately 28:1 across the dataset.

The video training data is sourced from SDFVD 2.0 [11], an enhanced version of the original Small-scale Deepfake Forgery Video Dataset. The SDFVD 2.0 consists of 456 real and 471 deepfake videos, each with a duration of 4-5 seconds at 720p resolution. Augmentation techniques include horizontal flip, rotation, shear, brightness, contrast correction, Gaussian noise addition and spatially rescaling.

The audio training data were obtained from the Logical Access partition of ASVspoof 2019 [4]. This subset comprised 2,580 real utterances and 22,800 synthetic utterances generated by seventeen text-to-speech and voice conversion systems. The development partition uses different speakers and synthesis systems for validation during model training.

TABLE III
Dataset Summary (= balanced sampling applied)*

Dataset	Type	Real	Fake	Used for
Mendeley 2025	Images	3,744*	3,744*	Image model training
SDFVD 2.0	Videos	456	471	Video model training
ASVspoof 2019	Audio	2,580*	2,580*	Audio model training

V. IMPLEMENTATION

The three models were implemented using PyTorch version 2.2, while the torchvision library loads pretrained EfficientNet-B0 weights. Audio processing utilized the Librosa 0.10.1, while visual processing employed OpenCV 4.9. The web interface was developed using Streamlit 1.32. All training was performed on CPU hardware. Overall, the project consists of three model files, three utility files and one application file, approximately 1,200 lines of Python code.

Each of the four distinct application operating modes follows a dedicated processing pipeline. Image detection loads the image model, runs inference and invokes Grad-CAM function for interpretability. Video Detection, the video model is loaded, frames are extracted, each frame is analyzed, and the results are aggregated. During audio detection, the audio model is loaded and converted into Mel spectrogram format before executing inference. Finally, the Cross-Modal Analysis function executes all three models on the input video and subsequently applies lip synchronization and consistency scoring tools to generate a forensic report.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05, May 2026
Received: 02-04-2026 Accepted: 07-05-2026 Published: 15-05-2026

VI. Results and Discussions

A. Model Performance

TABLE IV
Performance Evaluation Metrics

Modality	Accuracy	Precision	Recall	F1 Score
Image	98.17%	99.34%	96.98%	98.15%
Video	98.40%	97.87%	98.95%	98.41%
Audio	99.86%	99.73%	99.10%	99.86%

B. Discussion

The image model achieved 98.17% evaluation accuracy after 20 epochs. The approximately 2% gap between training and evaluation accuracy represents an acceptable generalization gap and indicates effective regularization. Earlier models trained on an imbalanced dataset without sampling rapidly achieved higher accuracy, but incorrectly classified all inputs as fake, including real images. The final model was able to recognize both classes accurately by introducing balanced sampling and an additional dropout layer.

The audio model produced the highest overall performance among the three pipelines and achieved 99.86% accuracy on a balanced test set consisting of 5,096 utterances from the ASVspoof 2019 development partition. This demonstrates that the model's ability to distinguish between real and synthetic speech effectively. Additionally, the results are consistent with prior ASVspoof research, indicating that CNN-based models using Mel spectrograms performed effectively for spoof detection under a balanced dataset. Initially, the class imbalance of 8.84:1 in the training data negatively impacted earlier trials without balanced samples. Therefore, equal class sampling is an essential component of the training process.

The video model achieved an accuracy score of 98.40% when evaluated on a balanced set of 4,560 frame samples consisting of 2,280 real and 2,280 fake, sourced from SDFVD 2.0. The corresponding F1 score was 98.41%, while the precision and recall values were 97.87% and 98.95%, respectively. Performance substantially improved the earlier evaluation accuracy of 82%, demonstrating the effectiveness of the frame sampling strategy. Extracting 10 uniformly spaced frames from each of the 927 yielded 9270 training samples. Earlier versions of the model trained using only video-level counts produced error rates of 39.3%. Another major contributor to high performance was obtained through partial fine-tuning, wherein the final blocks were trained while the earlier pretrained layers remained frozen. This preserved the generic visual features and enhanced generalization

The SDFVD 2.0 algorithm was evaluated on face-swapped videos from SDFVD 2.0 containing original audio and manipulated facial imagery. In the majority of cases, the model correctly classified these sample face-swapped attacks by combining the outputs of the pipelines. The audio model detected the speech as real, while the visual model detected the frames as fake. Neither of the two models can reliably detect the full forgery independently. An audio-exclusive pipeline cannot detect face manipulation, while a visual-exclusive model cannot validate audio authenticity. The integration of models with lip-synchronization analysis enabled accurate classification of the manipulation type, demonstrating the effectiveness of multimodal detection.

VII. Conclusion

This paper presents DeepGuard, a browser-based system for detecting manipulated images, videos, and audio content. The audio pipeline utilizes Mel spectrograms and trains a custom four-block convolutional neural network for classification. Additionally, image and video models use an enhanced version of the EfficientNet-B0 network to predict classes, while the cross-modal consistency module fuses predictions from all three modalities and evaluates lip sync consistency alignment. Experimental results show the accuracy of 98.17% for image detection, 98.40% for video detection and 99.86% for audio detection. The cross-modal module in SDFVD 2.0 successfully identifies face-swap attacks. With Grad-CAM explanations and the automated forensic reports, the framework is suitable for forensic and investigative applications.

VIII. Future Scope

Several future extensions can improve this research. Although the video model was solely trained on SDFVD 2.0, it produced a 98.40% frame-level accuracy score. Thus, evaluating the framework on diverse datasets such as Celeb-DF v2 and FaceForensics++ would provide deeper insights into the ability of the model to generalize to unseen manipulation methods. The current lip synchronization module relies on traditional OpenCV Haar cascade detection, which is sensitive to pose variation and lighting conditions. Replacing it with the MediaPipe Face Mesh application could improve the facial landmark localization and synchronizing mouth motion tracking. Moreover, the cross-modal module currently generates a single output per video. Incorporating the ability to identify and indicate instances of time discrepancies would enhance long video analysis.

References

[1] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... & Etzioni, O. (2025). Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*.

[2] 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).

[3] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).

[4] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... & Lee, K. A. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*.

[5] Jung, J. W., Heo, H. S., Tak, H., Shim, H. J., Chung, J. S., Lee, B. J., ... & Evans, N. (2022, May). Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6367-6371). IEEE.

[6] Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize?. *arXiv preprint arXiv:2203.16263*.

[7] Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

[8] Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., ... & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, *18*, 2015-2029.

[9] Cai, Z., Stefanov, K., Dhall, A., & Hayat, M. (2022, November). 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)* (pp. 1-10). IEEE.

[10] Islam, Md Raisul; Rakib, Md. Aminul Islam; Sahin Afridi, Arafat; Islam, Mohammad Monirul (2025), "Comprehensive Deepfake Detection Dataset: Real and Synthetic Frames from Roop and Akool AI Technologies", Mendeley Data, V1, doi: 10.17632/pdcp9mjy3z.1.

[11] Kaman, Shilpa; Makandar, Aziz (2025), "SDFVD2.0: Extension of Small Scale Deep Fake Video Dataset", Mendeley Data, V1, doi: 10.17632/zzb7jyy8w8.1

[12] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[13] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

[14] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion*, *64*, 131-148.

[15] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & Ling, Z. H. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*, 101114.