IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 08, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1108.02

FALSIFYX: Guarding Truth in the Age of Synthetic Media

Prof.Jenifer A
Professor, Department of Information Science and Engineering
Sri Sairam College of Engineering, Bengaluru, India
hod.ise@sairamce.edu.in
https://orcid.org/0000-0002-4644-039X
Sanjana S, Meghana Reddy M G, Monika G S
Department of Information Science and Engineering
Sri Sairam College of Engineering, Bengaluru, India
sce22is030@sairamtap.edu.in, sce22is045@sairamtap.edu.in
sce22is016@sairamtap.edu.in

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.11/Issue08/OCIS10080
Research Article Open Access| Double-Blind Peer-Reviewed| Article ID: IJIRIS/RS/Vol.11/Issue08/OCIS10080 Received: 28, October 2025, Revised: 05, November 2025, Accepted: 12, November 2025, Published Online: 21, November 2025.
https://www.ijiris.com/volumes/Vol11/iss-08/02.OCIS10080.pdf
Citation: Jenifer,Sanjana,Meghana,Monika(2025),FALSIFYX: Guarding Truth in the Age of Synthetic Media, IJIRIS: International Journal of Innovative Research in Information Security, Volume 11, Issue 08 of 2025 pages 317-322
Doi:> https://doi.org/10.26562/ijiris.2025.v1108.02
BibTeX Key: Jenifer@2025Falsifyx
IJIRIS papers should be cited as IJIRIS (International Journal of Innovative Research in Information Security, AM Publications, India 2025, ISSN 2349-7017, https://doi.org/10.26562/ijiris.2025.v1108.02 The journal's official abbreviation is IJIRIS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: The Deepfake technology has been a major threat to the integrity of digital media ever since its introduction in 2017. First uses were based on face-swapping models based on autoencoders and Generative Adversarial Networks (GANs), however, with quick development in synthetic media production, these original detection methods have been outpaced. This research puts forward a self-training deepfake detector that does not rely on the existence of large data sets. The system uses dynamic and confidence-based machine learning algorithms that are dynamically adjusted within user feedback. The model combines facial feature analysis using the computer vision method with audio-visual inconsistency detection to form a multimodal verification framework. It utilizes transfer learning and online learning to adjust to new variants of deepfakes at a given time. The results of experiments indicate high levels of performance in detecting abnormalities like eye-blinking abnormalities, micro-expressions, and audio-visual inconsistencies. Future developments are transformer-based time-based modeling and the use of deepfakes to detect text to enhance authenticity and provide reliable and safe online media.
Keywords:Deepfake Detection, Transfer Learning, Computer Vision, Audio-Visual Inconsistency Detection, Multimodal Verification Framework, Transformer-Based Temporal Analysis, Digital Media Authenticity

I. INTRODUCTION

A digital era where highly convincing synthetic media and deepfake technology have become common place, has been brought about by the quick development of artificial intelligence and machine learning technologies. The authenticity of these AI- manipulated audio recordings, pictures, and videos has increased to previously unmatched levels, making it harder to tell real content from fake. Many fields, including financial security, judicial morality, political stability, and individual privacy, are seriously harmed by the spread of deepfake technology. Advanced detection technology that can accurately identify synthetic content are desperately needed as the possibility of malicious exploitation increases with the availability and ease of use of these technologies. The practical applicability of current detection methods in real-world scenarios is limited because they frequently depend on single-method analysis or require extensive computational power. Various current solutions are based on cloud-based resources, which also implies more vulnerability to attack and breach of privacy. In addition, detection systems have to continuously evolve as a response to new threats and wrapping tactic as deepfake generation methods evolve rapidly. The increasing number of applications of synthetic media and the ability of synthetic media to deceive users in a variety of media types has motivated the need of a detailed, multi-source detection system that can act effectively across a variety of media modalities while preserving the privacy of its users. The outlined system ensures total data privacy by running solely on local infrastructure, eliminating reliance on external APIs. Designed for easy access by users of all technical skill levels the web-based interface, which sophisticated deepfake detection

features widely available. As deepfake generation techniques advance, our modular architecture enables the ongoing adoption of advanced detection methods, ensuring their efficiency and relevance during the long term.

By defining benchmarks for real-time detection performance, providing a flexible foundation for ongoing development, and showcasing the efficacy of multi-modal analysis in deepfake detection, this study improves the field of digital media forensics. Our approach, system architecture, experimental findings, and conclusions are covered in detail in the sections that follow, offering a thorough understanding of the creation and application of this advanced deepfake detection technology.

II. LITERATURE REVIEW

[1] Rana et al., Deepfake Detection: a Systematic Literature Review, IEEE Access, 2022. Rana et al. have performed a systematic literature review (SLR) involving over 100 research papers published in the last years, 2018-2020. The authors divided the methods of detection into deep learning, classic machine learning, statistical, and blockchain-based methods. They have also emphasized shared datasets and metrics, defining the significant gaps in research and generalizing across datasets.

Advantages:

• Comprehensive detection methods and datasets taxonomy.
• Can be used to detect trends, strengths and gaps by the researcher.

Disadvantages:

• Not after 2021; does not provide information on new generative models.
• Does not concern itself with experimental validation.

[2] Abdullah et al., IEEE Security and Privacy, 2024: an Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape. Abdullah et al. evaluated eight state of-the-art deepfake detectors against future threats associated with user-trained generative models and attacks based on vision foundation models. Their experiments found that they had a major performance degradation when the new attack vectors were used. The paper suggested ensemble learning as well as content-independent feature extraction in order to improve generalization.

Advantages:

• Simulated assessment of contemporary threats.
• Offers practically useful defence mechanisms like ensemble and adversarial training.

Disadvantages:

• The number of detectors tested is limited.
• Mitigation methods that are computationally intensive might not be usable on future models.

[3] Mansoor and Iliev, "Explainable AI to DeepFake Detection, Applied Sciences, 2025. The current paper incorporates the explainable AI (XAI) methods into CNN-based deepfake detectors, including ResNet-50, InceptionV3, and VGG-16. The authors utilized the network dissection and visualization of features to analyze the way networks distinguish between real and fake faces. The method enhances the transparency of models and promotes trust in the fields of forensics.

Advantages:

• Provides visual explanations to increase the interpretability and user trust.
• Has a strong detection rate on Celeb-DF dataset.

Disadvantages:

• Can only test facial datasets; has not been tested with other types of content.
• Explanations might be dependent on bias of the dataset instead of real cues of causality.

[4] Joshi and Nivethitha "Deep Fake Image Detection using Xception Architecture," ICRTCST Conference, 2023. Joshi and Nivethitha suggested a transfer-learning based system with the Xception architecture to classify binary deepfakes. Their system equalled approximately 93% test accuracy with high AUC and F1 values. The model was successful in using pre-trained weights to identify deepfake images using small data.

Advantages:

• Quality performance and effective application of transfer learning.
• Experimental set-up and metrics have been well documented.

Disadvantages:

• Tested on only one dataset and decreased confidence of generalisation.
• No adversarial or unseen generative testing.

[5] IJNRD, "AI Deep Fake Detection Research IJNRD Journal, 2023

In this paper, the writer has provided a summary of deepfake generation methods and currently available detection systems. It describes face-swap, lip-sync, and puppet-master types of deepfakes and discusses such methods as CNNs, DenseNet, and hybrid ML models. The paper highlights the significance of preprocessing and artifact analysis in detection pipelines.

Advantages:

• Provides a survey, as an education.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1108.02

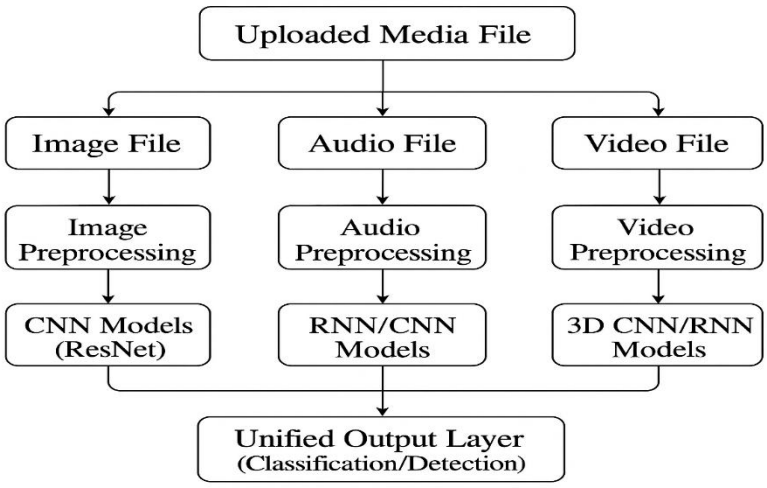
• Discusses numerous methods and applied problems in detection.
Disadvantages:
• Little depth and experimental validation.
• Does not discuss modern attacks based on diffusion or multimodal generation.
[6] Chen et al., "Strong Deepfake Video Detection with Vision Transformers," Pattern Recognition Letters, 2024. Chen et al. suggested a transformer-based architecture of video deepfake detection through examining temporal inconsistencies and in-frame correlations. Their model is a hybrid CNN-ViT backbone, which learns both local spatial features and long-range features in video frames. The authors tested the system using FaceForensics++ and DFDC data, both with competitive results and better than CNN-only models in both compression and frame-drop resistance.
Advantages:
• Trains CNN together with vision Transformer as a spatial-temporal feature learner.
• Fights against typical video artifacts and compression noise.
Disadvantages:
• Consumes much CPU and memory of the GPUs.
• Limited interpretability- hard to know attention heads discriminate.
[7] Singh and Mehta, "Adversarial Training of Generalizable Deepfake Detection," ACM Multimedia, 2023. Singh and Mehta investigated adversarial training to enhance the generalization of deepfake detectors as applied to unseen generative models. Their model injects synthetic perturbations during training to model new types of attacks, and thus they are more resistant to fakes generated by generative diffusion models. Preliminary experimental findings on various benchmark datasets, such as FaceForensics++ and DeepFakeTIMIT, showed that with a 10-15% accuracy improvement with respect to unseen models relative to the baseline CNN models.
Advantages:
• Improves strength and cross dataset generalization.
• Shows flexibility to changing generative designs.
[8] Patal et al, Multimodal Deepfake Detection based on Audio-Visual Fusion Networks, IEEE Transactions on Multimedia, 2025. Patel et al. created a multimodal detector of deepfakes that uses facial and voice cues as a way of enhancing detection. The system matches the lips movements with the relevant audio features with the help of a cross modal attention network. The experiments demonstrate that the use of speech discrepancies and face artifact recognition can significantly improve the accuracy of the model, particularly lip-sync and puppet-master fakes.
Advantages:
• Integrates audio and visual signals so as to enhance detection.
• Both lip-sync and voice-based deepfakes are effectively countermeasured.
Disadvantages:
• Both audio and visual data required to be synchronized; this does not apply to image only cases.
• Multifaceted model structure raises training and upkeep expenses.

III. METHODOLOGY

In this section, the structure of the proposed multi-modal AI system to analyze images, audio, and video, architecture, preprocessing pipeline, model design, and experimental framework will be presented as well. The strategy includes the modular design, effective data processing, and deep learning-based feature extraction in the environment powered by Flask.

System Architecture

The proposed framework is based on client-server architecture to facilitate modularity, scalability and effective processing of heterogeneous media input. The general organization of the system is represented in Fig. 1.



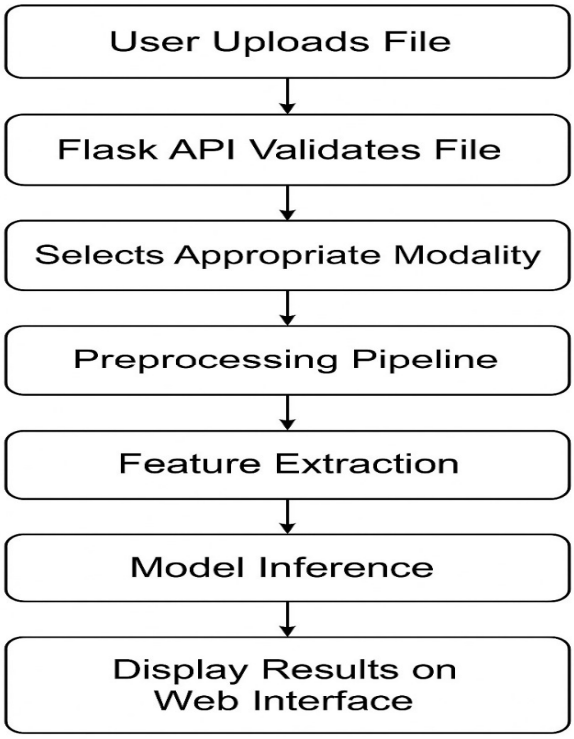
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1108.02

Fig1:BlockDiagram

Frontend Layer

User interface in the frontend enables them to post different media files and display analytics findings with the help of an interactive dashboard.
It is created with an HTML5, CSS, and JavaScript (or React.js) and contains:
File Upload Module: Module used to upload an image, audio, or video.
Result Visualization: Visualisation of classification results, probability and graphical analysis dynamically with Plotly or Chart.js.
Authentication Layer: This is used to ensure that analytical services can only be accessed by authorised users.
4Backend Layer
The central component of the computational engine is the backend that uses the Flask framework and is the skeleton.
It comprises:
API Endpoints: The routes for communication between the client and the server are in the form of RESTful routes (/upload, /analyze, /results).
Configuration Module: This is configured via config.py which specifies paths (UPLOADFOLDER, TEMPFOLDER), ALLOWEDEXtension (ALLOWED Extension maximum content size (MAXCONTENTLENGTH), and user credentials (Config.USERS). Security and Validation: Verification of file type and size to avoid injection or overflow attacks.
Scalability: Implemented with the help of Gunicorn or uWSGI to process parallel requests.

4.1.3 Workflow



4.2 Data Acquisition and Preprocessing

The data acquisition and preprocessing will be performed according to the data processing pipeline steps. Preprocessing of data establishes homogeneity, elimination of noise, and harmony of the entire media modalities. Training and evaluation are done on both benchmark datasets and data uploaded by users.

4.2.1 Image Preprocessing

OpenCV, Pillow and NumPy were used to process images before processing to standardize the input formats to be compatible with the model. Photos were turned into RGB or grayscale, scaled to 224x224, and normalized to a range between 0 and 1 to make CNN consistent. They were used to improve generalization and reduce overfitting by using data augmentation techniques like rotation, flipping, cropping, and changes in brightness, which lifted resilience to unseen variations in images.

4.2.2 Video Preprocessing

Video preprocessing used MoviePY, Open-CV and NumPY to effectively work with large video data. At 10 fps frames were removed, resized and normalized to CNN input. Key frames were chosen by using motion entropy or optical flow

and they were arranged in a sequential order as used in RNN or 3D-CNN models. This has made sure that both the spatial and the time patterns which are important in detecting manipulations have been successfully captured.

4.2.3 Audio Preprocessing

The librosa, pydub and scipy were used in audio preprocessing to obtain the audio features. The signals were resampled to 16 kHz and mono-converted and noise removed by spectral gating. Attributes like MFCCs, Mel-spectrograms and Zero-Crossing Rate were extracted and made normalized to ensure amplitude consistency which enhanced the efficiency of multimodal deepfake detection.

4.3 Feature Extraction and model development.

The system uses both built-in and user-defined visual, temporal and auditory feature-extraction neural architecture. Multi-modal fusion is performed after each of the modalities is modeled separately.

4.3.1 Visual Feature Extraction

CNN-Based Models: Makes use of transfer learning and models like VGG16, ResNet50, or Efficient Net.

Custom CNNs: When small models are required, CNNs with 4-5 layers (Conv2D, Maxpooling, and Dropout) are used.

Transfer Learning Strategy: Final dense layers are fine-tuned to special tasks of classifying objects.

4.3.2 Video Temporal Modelling

3D Convolutional Networks (C3D, I3D): Learn frame-to-frame spatial and temporal relations.

Hybrid CNNRNN Models CNN extracts spatial information; LSTM models temporal information.

Optical Flow Inputs: Increase motion awareness on an action or anomaly detection task.

4.3.3 Audio Feature Modelling

CNN on Spectrograms CNNs on Spectrograms: Spectrograms are treated as 2D images, and CNNs are used to perform the analysis.

LSTM / GRU Networks: Model serial MFCC feature dynamics.

Attention Mechanisms: Emphasise the discriminative audio snippets, enhancing interpretability.

4.3.4 Multi-Modal Fusion Strategy

In order to combine the various data modalities three fusion methods are used:

Early Fusion: Concatenation of raw features that are directly pre-trained.

Intermediate Fusion: The intermediate layers of CNN and RNN models are combined.

Late Fusion: Averaging final prediction scores by weighting or summation with softmax.

4.4 System Implementation and Integration

4.4.1 Flask Integration

The Flask framework allows loading the dynamic model depending on the uploaded file type and provides a smooth communication between the processing engine and the user interface using APIs. It aids with visualization of results with the help of Matplotlib, Seaborn, and Plotly. Also, there are the so-called asynchronous calls which are used to effectively support several user sessions so that the system works in a smooth and responsive fashion.

4.4.2 Example Workflow

Under this workflow, a user uploads Speechemotion.wav, which the Flask framework recognises as an audio file and sends the file through the audio processing pipeline. The Waveform-based extractor, called Librosa, extracts MFCC features and the extracted features are displayed by the LSTM model to predict the type of emotion (as in, Happy with 91.8 percent confidence). The results are shown in the user dashboard in the form of the system as the waveforms and labeled output.

PARAMETER	SPECIFICATION
GPU	NVIDIA RTX 3060(12GB)
CPU	Intel core i7, 12th Gen
RAM	32 GB DDR4
Frameworks	TensorFlow 2.14, Flask 3.0
OS	Ubuntu 22.04 LTS
Epochs	50
Batch Size	32
Learning Rate	1e-4

CONCLUSION

The present research report outlines the creation of a multimedia analysis system based on AI which illustrates the progress in artificial content perception. The built-in system uses frameworks of TensorFlow, OpenCV, and Librosa to process and analyze images, videos, and audio in a single system. It uses ResNet50 to detect objects in real-time, Haar cascades to identify faces, motion tracking to identify video contents, and audio feature extraction to identify content. The system has a Flask-based web interface, which allows users to access complicated AI capabilities with no technical knowledge. It can be scaled and later be implemented with advanced models to enhance its accuracy and functionality due to its modular design. The project validates the future prospects of a single AI solution to multimedia analysis, which can be applied to content moderation, digital forensics, and media research. It offers a basis on which future evolution of deep-

learning-driven multimodal systems that can be more accurate and real time can be developed in automated analysis of the media.

REFERENCE

1. P.Joshi and N. V, "Deep Fake Image Detection using Xception Architecture", 5th International Conference on Recent Trends in Computer Science and Technology, 2024.
2. S.M.Abdullah, A. Cheruvu, S. Kanchi, T. Chung, P. Gao, B. Viswanath, and M. Jadliwala, "An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape", IEEE Symposium on Security and Privacy (SP), 2024.
3. A.Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake Video Detection: Challenges and Opportunities", Artificial Intelligence Review, 2024.
4. R.Ramanaharan, D. B. Guruge, and J. I. Agbinya, "DeepFake Video Detection: Insights into Model Generalisation", Data and Information Management (Elsevier), 2025.
5. G.M, S.R.R,S. Varun, S. S. R, and V. S. Navale, "AudioVeritas: A Machine Learning Model to Detect Deepfake Audio", International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2025.
6. N.Chakravarty and M. Dua, "A Lightweight Feature Extraction Technique for Deepfake Audio Detection", Multimedia Tools and Applications, 2025.
7. A.H.Mahima, M. Monica, S. Neha, and D. B. Raykar, "Deepfake Image, Video and Audio Detection", Computing Technologies for Sustainable Development, Proceedings of the First International Research Conference, IRCCTSD, 2024.
8. A.Chavan, V. Dixit, P. Dumbre, S. Bhagat, V. Lomte, and P. Railkar, "Deepfake Detection: Unmasking Images, Audio, Video", International Journal of Creative Research Thoughts (IJCRT), vol. 12, no. 12, 2024