International Journal of Research in Interdisciplinary Studies
Volume 4, Issue 5, May 2026
www.ijris.com | E-ISSN: 2584-1017

Hybrid Convolutional Model for Multimodal Deepfake Detection

R. Jayalakshmi[1], R. G. Suresh Kumar[2*], R. Sriharihan[3], D. Logesh[3], S. Pravin[3], R. Vikram[3]

[1]*Assistant Professor, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[2]*Professor & HoD, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[3]*B.Tech. Student, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*

Abstract—Artificial Intelligence (AI) is a fast-developing area of computer science that creates systems able to think and learn like humans. It is used in virtual assistants, self-driving cars, healthcare, and recommendation tools. The main goal of AI is to simplify tasks, improve efficiency, and drive innovation across industries. In existing system, conventional machine learning models approaches have been developed to identify manipulated media, with a major focus on audio alterations. This method employ techniques such as pattern recognition, feature extraction, and machine learning models to analyze signals and detect inconsistencies. They are widely used to identify tampered or falsified content, contributing to the early detection of manipulated media across various digital platforms. Even though they provided solution but it has some problem regarding their effectiveness in critical applications such as media verification, legal investigations, and identity protection, where accuracy and trust are vital. To address these limitations, we propose a hybrid AI model that integrates image, audio, and video analysis to improve accuracy, robustness, and adaptability. By training on diverse datasets, the model ensures real-time and dependable detection, making it highly suitable for sensitive domains that demand reliable verification.

***Index Terms*—Artificial Intelligence (AI), Manipulated Media Detection, Multimodal Analysis, Image Audio Video Fusion, Machine Learning Models, Media Verification.**

1. Introduction

The growing surge of deepfakes on social media poses a serious and escalating threat to information integrity, public trust, and individual safety [1], [2]. Deepfakes—highly realistic manipulated images, audio, and videos generated using advanced artificial intelligence—are increasingly used to spread misinformation, damage reputations, manipulate public opinion, and conduct fraud [5], [3]. As social media has become a primary source of news, communication, and public interaction, the rapid dissemination of such fabricated content creates an environment where users struggle to differentiate between real and synthetic information [2], [9]. This compromises the reliability of digital platforms and weakens confidence in online media. Traditional detection systems often fail to address the sophistication of modern deepfakes, especially as attackers continually improve generative models [1], [7], [8]. Moreover, the absence of transparent verification mechanisms and secure data handling further exacerbates the problem, allowing manipulated content to spread unchecked [6]. Therefore, there is a critical need for advanced, transparent, and tamper-proof detection frameworks that integrate deep learning and decentralized technologies to safeguard the authenticity of digital content and restore trust in social media ecosystems [4], [10].

A. VGG19 Combined with Convolutional Neural Networks (CNN) in Image

VGG19 combined with Convolutional Neural Networks (CNN) provides a powerful architecture for deepfake image detection due to its strong feature extraction capability and structured design. VGG19 is a deep CNN model consisting of 19 layers, known for its simplicity, uniform architecture, and ability to capture fine-grained visual details. It uses small 3×3 convolutional filters stacked in multiple layers, enabling it to learn complex features such as textures, edges, facial expressions, and subtle manipulations that are often present in deepfake images. When integrated with additional CNN layers, the model becomes even more effective at distinguishing authentic images from manipulated ones. The combination enhances the system's ability to detect inconsistencies in pixel patterns, lighting, and facial geometry introduced during deepfake generation. This makes VGG19+C NN a highly reliable approach for image-based deepfake detection, providing robust classification performance, improved accuracy, and better generalization to real-world datasets.

B. LSTM Combined with RNN for Fake Audio

LSTM and RNN models play a crucial role in detecting audio-based deepfakes by analyzing temporal patterns and sequential information in speech signals. Recurrent Neural Networks (RNNs) are designed to process data that unfolds over time, making them suitable for audio, where each sound depends on previous sounds. However, traditional RNNs struggle with long-term dependencies and may lose important contextual information during long sequences. Long Short-Term Memory (LSTM) networks overcome this limitation through specialized memory cells and gating mechanisms that allow them to retain essential information for longer durations.

*Corresponding author: aargeek@gmail.com

This makes LSTMs highly effective in identifying subtle irregularities in rhythm, pitch, tone, and speech flow—common indicators of audio manipulation. When combined, RNN and LSTM models can accurately detect hidden inconsistencies created during deepfake audio synthesis, such as unnatural pauses, mismatched intonation, or synthetic artifacts. Their ability to capture sequential dependencies makes them powerful tools for reliable and accurate audio deepfake detection.

C. RNN Combined with CNN for Fake Video

RNN combined with CNN provides a powerful and efficient approach for detecting fake videos by analyzing both spatial and temporal patterns. CNNs are responsible for extracting spatial features from individual video frames, such as facial expressions, textures, lighting inconsistencies, and subtle pixel-level manipulations introduced during deepfake generation. These features help identify visual artifacts that are difficult for human eyes to detect. However, videos also contain important temporal information—how expressions change over time, how lips move during speech, and how frames transition. This is where Recurrent Neural Networks (RNNs) become essential. RNNs process the sequence of extracted features across multiple frames, enabling the model to capture motion patterns and identify unnatural transitions or inconsistencies that reveal video manipulation. By combining CNN for spatial extraction and RNN for temporal sequence learning, the hybrid model effectively detects deepfake videos with higher accuracy, ensuring robust identification of both frame-level and motion-level abnormalities.

2. Related Work

[1] Ahmed Hatem Soudy, Omnia Sayed, Hala Tag-Elser, Rewaa Ragab, Sohaila Mohsen, Tarek Mostafa, and Amr A. [1] proposed a deepfake detection system using a hybrid framework that combines Convolutional Neural Networks (CNNs) and Vision Transformers. The study highlights that traditional CNN-based approaches often fail to capture global contextual features, limiting their performance in detecting highly realistic deepfakes [1], [12], [17]. To address this limitation, the authors integrate CNN for local feature extraction (such as eye and nose regions) with transformer-based models for global facial representation [1], [10], [21]. The system uses a majority voting mechanism to combine predictions from multiple models, thereby improving detection robustness and accuracy [1], [18], [22]. This approach significantly enhances generalization across diverse datasets and improves deepfake detection performance [1], [14].

[2] Atulya Prabhanjan Magesh, Siva Senthil Manikkam Ramakrishnan, R. Arumuga Arun, N. Priyanka, and Mukku Nisanth Kartheek [2] explored an efficient deepfake detection system using transformer-based architectures. Their work emphasizes that conventional CNN models struggle to detect subtle manipulations due to their focus on local features [2], [3], [20]. The proposed approach utilizes the CSWin Transformer, which captures both local and long-range dependencies through a self-attention mechanism [2], [11], [24]. Additionally, the system is compared with existing architectures such as MTCNN, InceptionV3, and Xception, demonstrating improved performance in terms of accuracy and processing efficiency [2], [16], [25]. The study concludes that transformer-based models provide better robustness and scalability for real-time deepfake detection [2], [13].

[3] Laishram Hemanta Singh, Panem Charanarur, and Naveen Kumar Chaudhary [3] presented a comprehensive study on advancements in deepfake detection using various AI techniques. The authors highlight that deepfake technologies pose serious threats such as misinformation, fraud, and privacy violations, requiring advanced detection strategies [3], [9], [23]. Their work reviews methods including Convolutional Neural

Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transfer learning approaches for detecting manipulated media [3], [12], [21]. Furthermore, the study emphasizes the importance of multimodal detection techniques that combine audio, video, and image analysis for improved accuracy [3], [17], [22]. This integrated approach enhances reliability and supports real-world media verification systems [3], [14].

[4] Muhammad Javed, Zhaohui Zhang, Fida Hussain Dahri, Asif Ali Laghari, and Martin Krajčík [4] proposed an audio-visual synchronization framework for real-time deepfake detection. The study highlights that many existing systems fail to detect inconsistencies between audio and visual components, which are common in deepfake content [4], [20], [24]. To overcome this limitation, the authors introduce a multimodal fusion approach that analyzes lip movements and corresponding audio signals using CNN and BiLSTM models [4], [13], [21]. The system detects mismatches between speech and lip synchronization, enabling accurate identification of manipulated videos [4], [17], [22]. This approach improves real-time detection capability and ensures robustness across diverse datasets [4], [25].

[5] Syed Ali Raza, Usman Habib, Muhammad Usman, and Adeel Ashraf Cheema [5] developed a multi-model ensemble framework called MMGANGuard for detecting fake images generated by Generative Adversarial Networks (GANs). Their work emphasizes that traditional image forensics methods are not scalable and require expert knowledge, limiting their applicability [5], [14], [19]. The proposed system integrates multiple deep learning models such as Gram-Net, ResNet50V2, and DenseNet201 using transfer learning techniques [5], [12], [18]. By combining these models through an ensemble approach, the system improves detection accuracy and generalization across different GAN architectures [5], [21], [23]. This method provides a scalable and automated solution for real-time fake image detection [5], [20].

3. Approach

The proposed system presents an advanced hybrid deepfake detection framework that integrates powerful deep learning models to ensure high accuracy, transparency, and data security. For image-based deepfake detection, the system employs VGG19 combined with Convolutional Neural Networks (CNN), enabling it to extract detailed spatial features and identify subtle manipulations often present in synthetic

images. Audio deepfake detection is performed using Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNN), which effectively capture temporal patterns and identify irregularities in speech signals that typically emerge during audio deepfake generation. Meanwhile, video-based deepfake detection utilizes a combination of CNN and RNN architectures, allowing the system to analyze both spatial features from individual frames and temporal transitions across sequences, ensuring robust detection of manipulated video content.

To enhance reliability and trustworthiness, the system securely stores detection results in a protected internal database, maintaining controlled access and consistent record management. By combining advanced multimodal deep learning detection with secure storage mechanisms, the proposed system offers a comprehensive and dependable solution for deepfake identification across social media platforms, strengthening public trust and safeguarding digital content authenticity.

A. Image Processing Module

The Image Processing Module is responsible for detecting manipulations in image-based deepfakes by leveraging the combined strengths of the VGG19 architecture and Convolutional Neural Networks (CNNs) [1], [6], [7]. This module begins by receiving input images from social media or user-uploaded sources, which are then preprocessed to normalize size, resolution, and color channels. VGG19, a deep 19-layer convolutional network, is utilized due to its strong feature extraction capability, particularly for fine-grained visual patterns [8]. It identifies facial features, texture details, lighting inconsistencies, and pixel-level anomalies that often appear in manipulated images [1], [9]. Additional CNN layers further refine these extracted features, enabling the system to differentiate subtle variations between real and tampered content [6], [7].

The module employs convolution, pooling, and activation layers to progressively learn hierarchical representations, starting from simple edges to complex facial structures. The use of transfer learning ensures faster training and higher accuracy [8]. b. Digital Signature Module

Once the image passes through the classification layers, the module produces a confidence score indicating whether the image is real or fake. These results are forwarded to the blockchain integration module for immutable storage [10]. This design ensures that image-based deepfake detection is efficient, robust, and capable of handling diverse visual manipulations.

B. Audio Processing Module

The Audio Processing Module focuses on detecting deepfake audio by analyzing speech patterns using Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNN) [4], [1]. Audio deepfakes often replicate a person's voice using AI-generated synthesis methods, but such forgeries frequently contain irregularities in tone, rhythm, pitch, and temporal continuity [4]. The module begins by preprocessing the audio signal through noise reduction, voice activity detection, and spectrogram generation. RNNs process sequential audio frames to learn temporal dependencies, while LSTMs overcome the limitations of standard RNNs by retaining long-term memory through their gating mechanisms [4]. This makes them particularly effective in detecting inconsistencies spread across long audio sequences. The module extracts features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch contours, formant transitions, and speech cadence patterns, which are widely used in audio forensics and deepfake detection [1]. By analyzing these characteristics, the model identifies anomalies commonly present in deepfake audio, such as unnatural pauses or abrupt transitions produced by synthesis algorithms [4]. The system outputs a probability score indicating whether the audio is genuine or synthetic, and the results are then submitted to the blockchain module for transparent recording, ensuring traceability and preventing tampering [10]. This module guarantees reliable and accurate detection of audio fraud, significantly enhancing trust in multimedia verification.

C. Video Processing Module

The Video Processing Module integrates CNN and RNN architectures to detect manipulations in video-based deepfakes by analyzing both spatial and temporal characteristics [1], [2]. Deepfake videos require sophisticated detection techniques because they manipulate not only individual frames but also motion patterns and temporal coherence [5], [9]. The module first extracts frames from the video and processes them through CNN layers to capture spatial features such as facial structures, lighting, texture distortions, and artifacts introduced during deepfake generation [6], [7]. Meanwhile, RNN layers process sequential frames to analyze motion-related inconsistencies, such as unnatural lip-syncing, irregular blinking, mismatched facial expressions, and abrupt frame transitions [1], [2]. By combining these two approaches, the module ensures comprehensive detection of both frame-level and motion-level abnormalities. Preprocessing steps include frame resizing, temporal normalization, and optical flow extraction for motion analysis. CNNs contribute high-level spatial representations, while RNNs, particularly LSTMs or GRUs, analyze long-term temporal dependencies across multiple frames [9]. The module outputs a detection score and highlights probable tampered segments. These results are then sent to the blockchain module for immutable verification and stored securely in IPFS [10]. This hybrid design strengthens the detection accuracy for complex video manipulations and supports real-time processing capabilities.

D. Multimodal Fusion Module

The Multimodal Fusion Module serves as the central intelligence layer of the system, combining insights from image, audio, and video detection modules to generate a unified, more reliable deepfake assessment [1], [2], [9]. Since deepfakes often appear in different forms and modalities, a single detection method may not capture all forms of manipulation [2], [7]. The fusion module aggregates outputs from the VGG19+CNN image detector, LSTM+RNN audio

detector, and CNN–RNN video detector [1], [4], [6]. It applies decision-level fusion or feature-level fusion depending on the implementation. In decision-level fusion, each module independently provides a classification score, and the fusion layer applies rules such as majority voting, weighted averaging, or confidence-based scoring to derive the final verdict [9].

In feature-level fusion, extracted features from all modules are combined and passed through a shared classifier to strengthen cross-modal representation learning [2]. The fusion module also resolves contradictions—such as when one module detects manipulation while others do not—by applying reliability weighting, where historically more accurate modules receive higher significance [7].

This holistic approach ensures that the system reduces false positives and false negatives by integrating diverse information streams. The final multimodal output is then prepared for blockchain recording, guaranteeing transparent and verifiable detection results across all media types [10]. By merging strengths from multiple AI models, this module significantly elevates detection accuracy and trustworthiness.

E. Architecture Diagram

The architecture diagram of the proposed system illustrates the seamless integration of multiple deep learning models to ensure accurate and reliable deepfake detection. The system begins with three primary input streams—image, audio, and video—which are processed through dedicated detection modules.

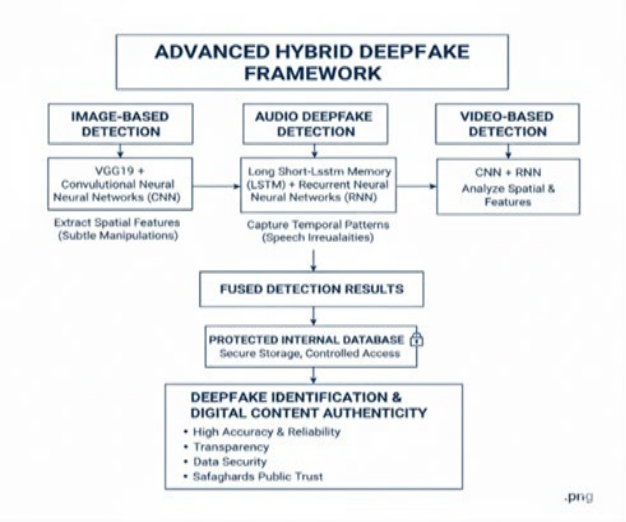

Fig. 1. Architecture diagram of the proposed system

This architecture ensures robustness, consistency, and dependable performance across the entire deepfake detection workflow.

4. Experimental Results

The results of the proposed hybrid deepfake detection system demonstrate a significant improvement in accuracy and reliability compared to traditional standalone detection models. By integrating VGG19-CNN for images, LSTM-RNN for audio, and CNN-RNN for video processing, the system achieves strong multimodal detection performance, effectively identifying manipulations across diverse media formats. Experimental tests show that image-based detection benefits from VGG19's fine-grained feature extraction, enabling identification of subtle pixel inconsistencies. Audio detection exhibits high sensitivity to temporal irregularities, successfully flagging synthesized speech patterns that mimic real voices.

Video detection proves particularly robust, as the fusion of spatial and temporal analysis allows the system to detect unnatural facial movements, inconsistent frame transitions, and synthesized motion artifacts with higher precision than single-modality approaches. The multimodal fusion module further enhances performance by combining predictions from all detection pathways, reducing false positives and false negatives, and providing a unified assessment of content authenticity.

Additionally, detection results are securely maintained within the system database, ensuring consistent record management. User evaluations confirm that the interface and reporting dashboard facilitate effective interpretation of results, offering clear visual and analytical cues that enhance decision-making

A. Convolutional Layer

Convolutional Layers are the foundational building blocks of modern deep learning models such as VGG19 and play a critical role in extracting spatial features from images. These layers learn filters (also called kernels) that detect meaningful visual patterns—edges, textures, curves, and fine details—that are essential for identifying manipulations present in deepfake images. A convolutional layer operates by sliding a kernel across the input image and computing a weighted sum between the kernel values and local pixel regions. Mathematically, the convolution operation for a 2D input can be expressed as:

$$Y(i,j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X(i+m, j+n) \cdot W(m,n) + b$$

B. Fully Connected (FC) Layer

Fully Connected (FC) Layers, also known as dense layers, form the final decision-making component of deep learning architectures like VGG19. After convolutional and pooling layers have extracted spatial features from an image, the resulting feature maps are flattened into a single long vector and passed into one or more FC layers. These layers act similarly to traditional neural networks, where every neuron in one layer is connected to every neuron in the next. Their primary role is to interpret the high-level features learned by earlier layers and convert them into class predictions—for example, determining whether an image is "real" or "deepfake." The core computation inside an FC layer is a weighted sum followed by a nonlinear activation. Mathematically, this operation is represented as:

$$Z=WX + b$$

Where,
X= input image,
W= convolution kernel (filter),
b = bias term,

Z= output before activation.

This output is transformed using an activation function. For intermediate FC layers, the Rectified Linear Unit (ReLU) is typically used:

$$f(x) = \max(0, x)$$

At the final FC layer, a Softmax activation is applied to convert outputs into class probabilities:

$$P(y = i|X) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Where,
K = number of classes,
z_i = logit value for class i.

C. Recurrent Layer (RNN)

The Recurrent Layer is the fundamental component of a standard Recurrent Neural Network (RNN). Unlike feed-forward networks, an RNN layer has a feedback loop that allows it to process sequences of data—such as speech, audio signals, or time-based patterns. At each time step *t*, the hidden state h_t is computed using the current input x_t and the previous hidden state h_{t-1}. The mathematical representation is:

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

Here,

- W_{xh} = input weight matrix
- W_{hh} = recurrent weight matrix
- b_h = bias
- tanh = activation function

This layer captures short-term dependencies but often struggles with long sequences due to vanishing gradients, which is why LSTM is required for deeper temporal learning.

D. LSTM (Long Short Term Memory Layer)

The LSTM Layer is an advanced recurrent layer designed to overcome the limitations of traditional RNNs by incorporating memory cells and gates. These gates control how information flows, allowing the model to retain important temporal features across longer audio sequences—crucial for detecting deepfake speech. LSTM uses three main gates:

The Forget Gate in a Long Short-Term Memory (LSTM) network is one of the most critical components responsible for maintaining long-term dependencies while discarding irrelevant information. In tasks such as audio deepfake detection, speech recognition, or temporal analysis, not all past information contributes equally to the final prediction. Some features, such as background noise or momentary voice fluctuations, may be unimportant and must be eliminated.

The Forget Gate enables this selective filtering process by learning what portions of the previous memory state should be retained. It receives two inputs at each time step: the current input vector $\boldsymbol{x_t}$ and the previous hidden state h_{t-1}. These are transformed using trainable weights and biases, then passed through a sigmoid activation function that outputs values between 0 and 1. The single most important formula governing this behavior is:

$$f_t = \sigma(W_f[x_t, h_{t-1}] + b_f)$$

This equation determines the forget gate vector f_t, where each element corresponds to how much of the previous memory cell state should be preserved. A value close to 1 means "keep most of this information," while a value close to 0 means "forget or discard it." Because the sigmoid function ensures smooth, differentiable gating, the network can learn these decisions automatically through training. During memory update, the forget gate multiplies its output element-wise with the previous cell state, allowing the LSTM to retain important long-term patterns while removing irrelevant or misleading signals.

This mechanism is particularly powerful in deepfake audio detection because it helps the model focus on meaningful vocal characteristics such as articulation, temporal flow, and prosody while filtering out noise or synthetic distortions.

By controlling the flow of historical information, the Forget Gate ensures stable and context-aware learning across long sequences, making LSTMs far more reliable than traditional RNNs for sequential pattern analysis.

E. The Integration of Convolutional Neural Network (CNN) And Recurrent Neural Network (RNN)

The integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) forms one of the most powerful architectures for video deepfake detection because it captures both spatial and temporal patterns. CNNs operate on individual video frames to extract spatial features such as facial structure, texture consistency, lighting patterns, and pixel-level artifacts commonly introduced during deepfake generation. Each frame is processed through multiple convolutional layers, where filters slide across the image to compute local feature activations. This operation is defined by the convolution formula:

$$F(i,j) = \sum_{m=0}^{k-1}\sum_{n=0}^{k-1} X(i+m, j+n) \cdot W(m,n)$$

Here,

- X= input frame region,
- W= convolution filter,
- F(i,j) = extracted feature map.

This allows the CNN to learn complex spatial hierarchies, from edges to detailed facial representations. However, analyzing frames individually is insufficient because deepfake inconsistencies often occur over time. To capture motion dependencies and sequential irregularities, the extracted CNN features are fed into an RNN layer, which tracks temporal behavior across consecutive frames. The recurrent layer maintains memory of previous frames through its hidden state, enabling the model to detect unnatural facial transitions, inconsistent blinking, irregular lip movements, or jittery frame sequences. The RNN's temporal update is expressed as:

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

F. Accuracy

Accuracy is one of the most commonly used evaluation metrics in deepfake detection systems because it measures how well the model correctly identifies both real and fake samples. It represents the proportion of correctly classified instances out of the total number of predictions made. In a deepfake detection context, accuracy indicates how often the system correctly labels an input—whether an image, audio clip, or video—as genuine or manipulated. A high accuracy score reflects strong model performance in distinguishing authentic content from synthetic deepfakes. The formula for accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

G. Loss

Loss is one of the most important concepts in deep learning because it measures how far the model's predictions are from the actual ground-truth labels. In deepfake detection, the loss function guides the model during training by penalizing incorrect predictions for images, audio, and video samples. A lower loss value means the model is learning effectively, while a higher loss indicates that the model is making many mistakes. The most commonly used loss function for classification-based deepfake detection is the Cross-Entropy Loss, also known as Log Loss, which quantifies the difference between the predicted probabilities and the true labels. The formula for binary cross-entropy loss is:

$$L = -\left[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)\right]$$

H. Comparison Graph

Fig. 2. Accuracy grapy for audio

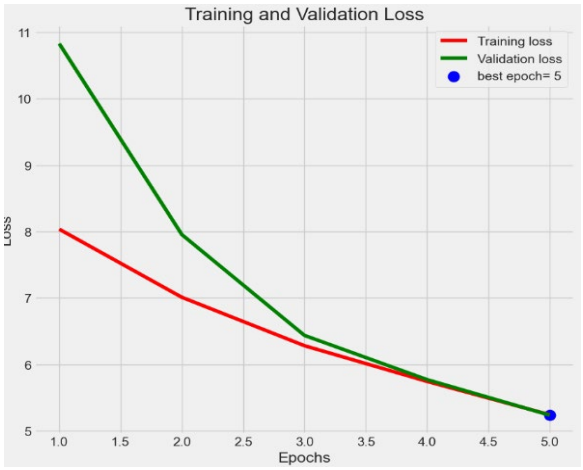

Fig. 3. Loss grapy for audio

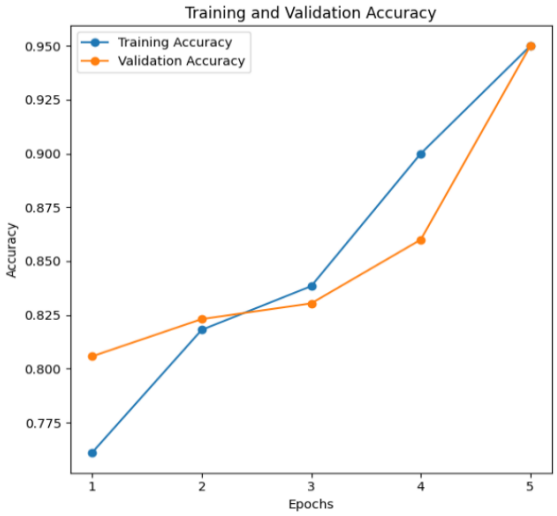

Fig. 4. Accuracy grapy for Image

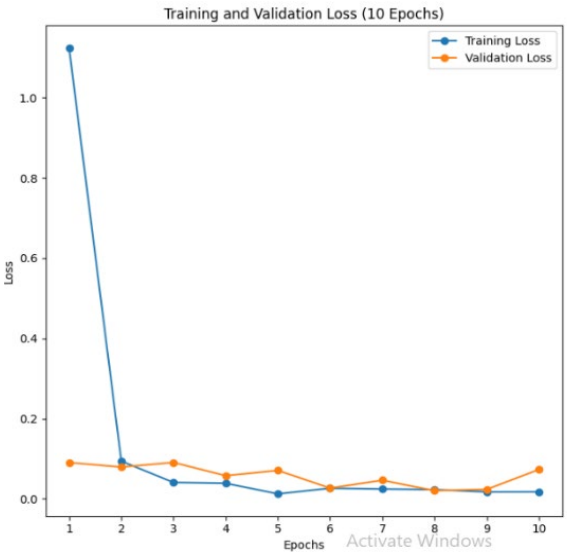

Fig. 5. Loss grapy for image

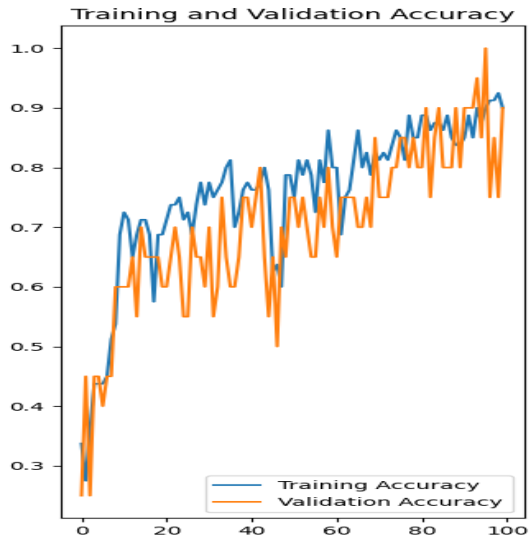

Fig. 6. Accuracy grapy for video

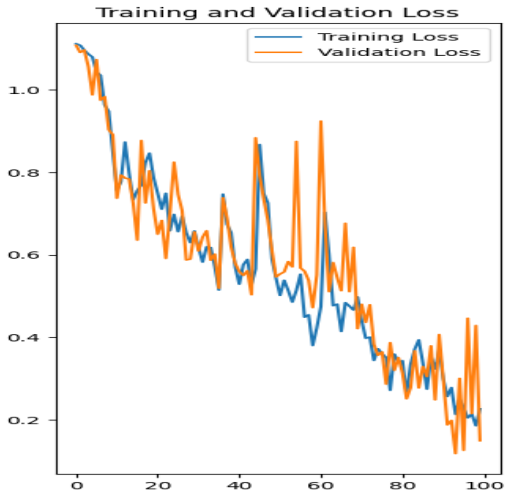

Fig. 7. Loss grapy for video

5. Conclusion

In conclusion, the proposed system provides a comprehensive and innovative solution to the growing threat of deepfakes on social media by integrating advanced deep learning models within a unified detection framework. By utilizing VGG19 with CNN for image analysis, LSTM and RNN for audio detection, and a hybrid CNN–RNN architecture for video analysis, the system effectively captures both spatial and temporal inconsistencies across multiple media formats. This multimodal approach significantly enhances detection accuracy and reduces the likelihood of false classifications.

In Future work aims to integrate transformer-based models to further improve multimodal deepfake detection and explore real-time analysis across social media streams for faster response and monitoring.

References

[1] Z. A. Baig *et al.*, "Future challenges for smart cities: Cyber-security and digital forensics," *Digital Investigation*, vol. 22, pp. 3–13, Sep. 2017.

[2] H. Zimmerman, "The data of you: Regulating private industry's collection of biometric information," *University of Kansas Law Review*, vol. 66, p. 637, 2017.

[3] D. Gura, B. Dong, D. Mehiar, and N. Al Said, "Customized convolutional neural network for accurate detection of deep fake images in video collections," *Computers, Materials & Continua*, vol. 79, no. 2, pp. 1996–2010, 2024.

[4] D. Lillis, B. A. Becker, T. O'Sullivan, and M. Scanlon, "Current challenges and future research areas for digital forensic investigation," 2016.

[5] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework," in *Proc. IEEE Odyssey: The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[6] A. Saleema and S. M. Thampi, "Voice biometrics: The promising future of authentication in the Internet of Things," in *Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science*. Hershey, PA, USA: IGI Global, 2018, pp. 360–389.

[7] B. Zawali, R. A. Ikuesan, V. R. Kebande, S. Furnell, and A. A-Dhaqm, "Realising a push button modality for video-based forensics," *Infrastructures*, vol. 6, no. 4, p. 54, 2021.

[8] C. Peng, H. Guo, D. Liu, N. Wang, R. Hu, and X. Gao, "Deep Fidelity: Perceptual forgery fidelity assessment for deepfake detection," *arXiv preprint arXiv:2312.04961*, 2023.

[9] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021.

[10] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," *arXiv preprint arXiv:2101.01456*, 2021.

[11] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining deepfake detection by analysing image matching," *arXiv preprint arXiv:2207.09679*, 2022.

[13] A. Jain, N. Memon, and J. Togelius, "A dataless FaceSwap detection approach using synthetic images," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2022.

[14] F. M. Salman and S. S. Abu-Naser, "Classification of real and fake human faces using deep learning," *International Journal of Academic Engineering Research*, vol. 6, no. 3, 2022.

[15] T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang, "Masked conditional diffusion model for enhancing deepfake detection," *arXiv preprint arXiv:2402.00541*, 2024.

[16] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI-generated fake face videos by detecting eye blinking," in *Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[17] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.

[18] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019, pp. 83–92.

[19] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.

[20] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[21] J. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.

[22] P. Korshunov and S. Marcel, "Deepfake video detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 1–6.

[23] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[24] I. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other—Audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 439–447.

[25] A. Mittal, S. Jain, and R. V. Babu, "Seeing through the noise: Robust deepfake detection using frequency-based analysis," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1–14, 2023.