ISSN: 2349-6002

Efficiently Identifying Fake Audio and Video Using Transfer Learning

Mrs. J. Veerendeswari[1], Mr. Mohamed Thoufiq K[2], Mr. Prabanjan P[3], Mr. Mohamed Musthafa A[4], Mr.Gowsik Roshan V[5]

[1]*Head of the Department, Information Technology, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*

[2,3,4,5]*UG, Information Technology, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*

doi.org/10.64643/IJIRTV12I11-201038-459

***Abstract*—The rapid rise of deepfakes on social media threatens information integrity, public trust, and personal reputation. Existing detection systems lack transparency due to reliance on centralized storage mechanisms. This proposed system introduces a hybrid deepfake detection framework that combines multiple deep learning models for comprehensive analysis. VGG19 with CNN is used for image analysis, while LSTM and RNN models handle audio deepfake detection by capturing temporal inconsistencies. Video tampering is identified through a combination of RNN and CNN architectures, ensuring accurate spatial–temporal analysis. The system securely stores detection results within an internal database to maintain data consistency and controlled access. By integrating multimodal detection with secure storage, the model improves accuracy, reliability, and trust in deepfake detection across social media platforms while maintaining data integrity and accessibility.**

***Index Terms*—Deepfake Detection, VGG19, CNN, LSTM, RNN, Ethereum, Multimedia Forensics. I.**

I. 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

ISSN: 2349-6002

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. 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.

II. LITERATURE SURVEY

[1] Vurimi Veera Venkata Naga Sai Vamsi, Sukanya S. Shet, Sodum Sai Mohan Reddy [1] This paper presents an innovative method for detecting Deepfake videos by combining ResNext, a robust Convolutional Neural Network (CNN) known for its proficiency in extracting intricate image features, with Long Short-Term Memory (LSTM) networks, which excel at analyzing temporal sequences in video data. This integrated approach offers a comprehensive analysis of Deepfake content: ResNext effectively captures spatial features from individual frames, identifying details and discrepancies that may indicate manipulation, while LSTM processes the temporal dynamics between frames, allowing the model to recognize patterns and changes over time. By synergizing these two powerful architectures, the proposed method significantly enhances the model's capacity to detect subtle alterations in Deepfake videos, leading to improved accuracy and greater robustness against the evolving sophistication of synthetic media. This dual-layer analysis not only addresses the challenges posed by static image features but also considers the context and progression of visual information, making it a promising advancement in the fight against Deepfake technology. [2]

Fakhar Abbas, Araz Taeihagh [2] This study investigates a variety of automated techniques aimed at both detecting and generating deepfakes in audio and images, offering a thorough review of existing frameworks, algorithms, and tools tailored for identifying synthetic media. By highlighting the effectiveness and limitations of these methods, the research provides valuable insights into the current capabilities and shortcomings of deepfake detection technologies. This comprehensive overview not only focuses on the technological aspects but also emphasizes the critical need for reliable detection mechanisms in the face of rapidly evolving deepfake

generation techniques, which pose significant risks to the authenticity of digital content. Furthermore, the study explores the application of these detection methods across different contexts, such as social media, journalism, and security, where the challenge of disinformation is particularly pressing. By analyzing how existing frameworks perform in these varied environments, the research aims to identify best practices and potential strategies for mitigating the risks associated with manipulated content. This analysis is crucial for developing effective countermeasures to combat the spread of deepfakes, ultimately contributing to the safeguarding of information integrity in an increasingly digital and interconnected world. Through this exploration, the study seeks to pave the way for improved approaches to deepfake detection, fostering a more informed and secure digital landscape. [3]

Ewout Nas, Roy de Kleijn [3] This study delves into a variety of automated techniques for both detecting and generating deepfakes in audio and images, offering a comprehensive review of the existing frameworks, algorithms, and tools developed to identify synthetic media. By assessing their effectiveness and limitations, the research sheds light on the current capabilities of deepfake detection technologies and the challenges they face in keeping pace with sophisticated manipulation methods. This detailed evaluation is crucial for understanding how these tools function and the conditions under which they may fail, thereby providing a foundation for further advancements in the field. In addition to examining the technological aspects, the study also considers the application of these detection methods across various contexts, including social media, journalism, and security. By addressing the growing challenge of disinformation, the research aims to highlight best practices and potential strategies for mitigating the risks associated with manipulated content in digital environments. This focus on practical application emphasizes the importance of effective detection mechanisms in real-world scenarios, ultimately contributing to the development of reliable countermeasures that can enhance the integrity of information and protect users from the harmful effects of deepfakes. Through this exploration, the study aspires to inform future research and the implementation of effective policies in combating digital misinformation. [4]

Mouna Rabhi, Spiridon Bakiras [4] The study reveals a critical vulnerability in leading audio deepfake classifiers, notably the Deep4SNet model, indicating that these systems are highly susceptible to adversarial attacks. These attacks involve the introduction of subtle perturbations to input data, which can cause classifiers to erroneously classify manipulated audio as authentic. Such vulnerabilities represent a significant threat to the reliability of audio deepfake detection technologies, as even minor alterations can drastically compromise the accuracy of these classifiers, undermining their effectiveness in real-world applications. These findings emphasize the urgent need for continued research and development of robust detection methods capable of resisting adversarial attacks. As deepfake technologies evolve and become more sophisticated, ensuring the integrity of audio verification processes is paramount in an increasingly deceptive digital environment. The study advocates for innovative strategies and enhanced training techniques that can fortify audio classifiers against such vulnerabilities, ultimately contributing to more reliable and secure audio verification systems. By addressing these challenges, the research aims to bolster trust in digital media and protect against the misuse of manipulated audio content. [5]

Michael Hameleers, Toni G.L.A. van der Meer, Tom Dobber [5] Hyper-realistic deepfakes that incorporate believable content manipulation often lead viewers to perceive them as credible, resulting in a tendency to accept fabricated media as authentic. This enhanced realism fosters a sense of trust among audiences, making it easier for them to be misled by manipulated content. As viewers encounter these convincing alterations, they may become less discerning and more vulnerable to the persuasive power of the deepfake, thereby amplifying the potential for misinformation to spread unchecked. The allure of realism in deepfakes underscores the importance of developing robust detection methods and raising awareness about the risks associated with such technologies. Conversely, deepfakes that feature less plausible content manipulation can have a paradoxical effect, significantly undermining the legitimacy of the individuals portrayed, such as politicians. When viewers detect discrepancies or implausibility's, even within less realistic contexts, it can erode their trust in the individual and the message being conveyed. This

erosion of trust can lead to reputational damage and a loss of credibility for the depicted figure, regardless of the intended impact of the deepfake. This dynamic illustrates the complex interplay between the perceived realism of deepfakes and their capacity to influence public perception. In light of this, fostering critical media literacy is essential in the age of synthetic media, empowering audiences to navigate and discern the authenticity of the content they encounter. [6]

M. Anil, K. Shiva, Y. Tulasi, G. Jayasri, B. Prudhvi Teja, E.K.M. Sai [6] Image recoloring is a sophisticated technique used to transfer colors or themes between images while ensuring that the changes remain imperceptible to the human eye. This process involves altering the color properties of an image in a way that preserves its original content and structure, making the transformation seamless and visually appealing. The technique can be particularly useful in various applications, such as enhancing artistic expression, adjusting images for aesthetic consistency, or preparing visual content for specific contexts, such as advertising or branding. The proposed network for image recoloring takes a novel approach by incorporating not only the original image but also two derived inputs that account for illumination consistency and inter-channel correlation.

By analyzing these factors, the network can output a probability indicating the likelihood that the image has been recolored. This design enables the model to effectively learn the relationships between colors in the original image and ensure that the recoloring process adheres to the natural lighting conditions and color dynamics of the scene. The resulting output enhances the model's ability to produce realistic recolored images while minimizing perceptible alterations, thereby maintaining the integrity of the original visual experience. [7]

Vandhana S, Vishnupriya J, Athira C, Saritha C, Reshma Mohan A [7] In today's digital landscape, the proliferation of fake images poses a significant threat to individuals and society at large, as people can easily be misled by manipulated visuals in their daily lives. This issue is exacerbated by the capabilities of Generative Adversarial Networks (GANs), which can create photo-realistic images from low-dimensional random noise. While these advanced technologies have revolutionized various fields, including image and video generation, they also facilitate the creation of misleading content that can be widely shared on social media platforms, leading to misinformation and other serious challenges. The rise of such deceptive media emphasizes the urgent need for effective and efficient image forgery detection mechanisms to identify and counteract the spread of fake images. Recent advancements in GANs have demonstrated their success in producing highly realistic images, which complicates the tasks of visual forensics and model attribution. One of the more concerning applications of GANs is imageto-image translation, which involves learning a mapping between source and target images, allowing for the generation of images that can easily deceive viewers. This paper provides a comprehensive overview of GANs, including their theoretical foundations and the challenges they present for detecting fake images generated by these networks. By discussing methods to identify GAN-generated content, particularly on social media, the study highlights the pressing need for vigilance and sophisticated detection techniques to mitigate the dangers posed by artificially generated images in an increasingly deceptive digital environment. [8]

Anushka Singh; Jyotsna Singh [8] the detection of fake images is essential for preserving the credibility of digital content, particularly in today's media landscape where misinformation can spread rapidly through social networks. As image forgery techniques become more prevalent and sophisticated, they pose a significant threat to the authenticity of online content. This paper introduces a deep learning-based approach for image forgery detection, utilizing Error Level Analysis (ELA) alongside Convolutional Neural Networks (CNNs) and a pretrained VGG-16 model. Through rigorous experimentation, the study compares the performance of the ELA-CNN model against the VGG-16 model, revealing that the ELA-CNN achieves an impressive accuracy rate of 99.87%, successfully identifying 99% of invisible image alterations, while the VGG16 model records a lower accuracy of 97.93% and a 75.87% validation rate. The findings underscore the importance of leveraging deep learning techniques for effective image forgery detection and highlight the potential implications of

these results for developing more reliable detection tools. The research not only sheds light on the comparative effectiveness of different deep learning algorithms in identifying manipulated images but also addresses the limitations of the current study and suggests avenues for future enhancements. By improving the precision and generalization capabilities of image forgery detection algorithms, this research aims to contribute significantly to the field, ultimately facilitating the creation of advanced tools that safeguard the integrity of digital content and mitigate the adverse effects of image manipulation in the digital realm. [9]

Md Shohel Rana; Mohammad Nur Nobi; Beddhu Murali; Andrew H. Sung [9] Deepfake technology can create convincing impersonations of individuals, posing significant risks such as identity theft and unauthorized access to sensitive information. As fake videos and audio are increasingly used in various forms of cyberattacks, including spear phishing and social engineering, it becomes crucial to implement robust detection mechanisms to prevent unauthorized access to critical data. The ability to identify and label manipulated content is essential for enabling individuals and organizations to take proactive measures against the spread of harmful misinformation. This not only helps protect the reputations and privacy of individuals but also plays a vital role in combatting fake news, fraud, and cyberbullying. Fortunately, advancements in deepfake detection tools are on the rise, signaling progress in this crucial area. Detecting deepfakes requires interdisciplinary collaboration among experts in computer science, artificial intelligence, psychology, and law. Deepfake detectors can analyze visible biometric indicators within videos, such as a person's pulse or voice characteristics produced by human vocal cords rather than synthetic sources. However, the tools developed to improve detection may also inadvertently aid in the creation of the next generation of deepfakes. Therefore, deepfake detection is a rapidly evolving field with significant societal implications. Ongoing collaboration among researchers, policymakers, and the public is essential to develop effective detection methods, address legal and ethical concerns, and raise awareness about the potential dangers of deepfakes, ultimately mitigating their harmful impacts.[10]

Sara Migliorini, Mauro Gambini, Alberto Belussi [10] the preservation and restoration of cultural heritage have gained significant importance due to their immense historical and touristic value. However, these activities require substantial financial resources, which cannot always be supported solely through public funding. Crowdfunding platforms have emerged as an effective way to raise funds for such projects, and several studies have explored their use for cultural heritage restoration. Despite their potential, concerns about transparency, reliability, and trust still limit donor participation. Blockchain technology offers a promising solution by ensuring immutability, traceability, and trustworthiness throughout the funding process.

However, systems that rely on cryptocurrencies often face resistance due to regulatory uncertainties and user skepticism. To overcome this, the proposed solution integrates blockchain with traditional crowdfunding platforms without using cryptocurrencies. It utilizes smart contracts and a decentralized application (dApp) to store information immutably on the blockchain. Donors can then independently verify the authenticity and flow of funds, enhancing trust and transparency in cultural heritage restoration projects.

III. PROPOSED SYSTEM

The proposed system presents an advanced hybrid deepfake detection framework that integrates powerful deep learning models to ensure high accuracy, transparency, and data security [1], [2], [9]. 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 [6], [7], [8]. 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 [4]. 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 [1], [2]. To enhance reliability and trustworthiness, the system securely

ISSN: 2349-6002

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

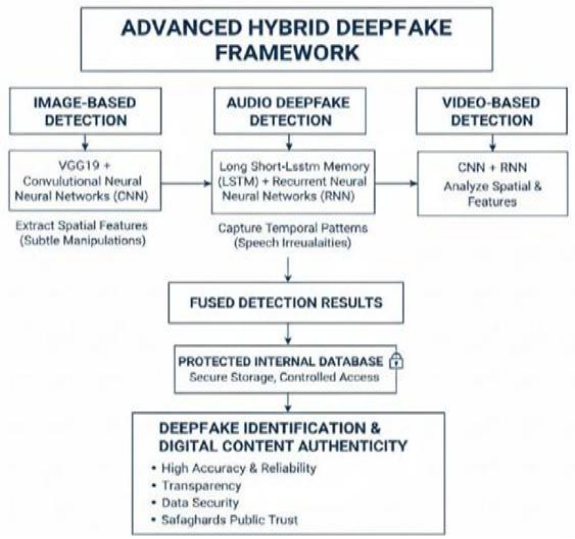

Fig 3.1 architecture diagram of the proposed solution for deepfake identification across social media platforms, strengthening public trust and safeguarding digital content authenticity [3], [5].
system

a. 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. Image data is fed into the VGG19+CNN model for spatial feature extraction, audio data is analyzed using LSTM and RNN networks to capture temporal inconsistencies in speech, and video data passes through a hybrid CNN–RNN pipeline to detect both spatial and temporal manipulations. After processing, the detection results from all modules are combined in a decision layer that generates the final authenticity prediction. The outcomes, along with relevant metadata, are securely stored in an internal database for record management and future verification. This architecture ensures robustness, consistency, and dependable performance across the entire deepfake detection workflow.

b. 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]. 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.

c. 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 feature 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.

d. 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.

e. 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.

IV. RESULT AND DISCUSSION

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

ISSN: 2349-6002

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 Layers

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:

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

Were,

X= input image, W= convolution kernel (filter), b = bias term, k= kernel size (typically 3×3).

Y(i,j) = output feature map.

This operation allows the layer to learn local patterns, a crucial property for deepfake detection, as fake images often contain subtle artifacts such as unnatural blending, inconsistent lighting, or smoothing distortions. After convolution, the result is passed through an activation function typically the Rectified Linear Unit (ReLU), defined as:

f(x) = max (0, x)

ReLU introduces non-linearity, enabling the network to learn complex patterns beyond simple linear transformations. Multiple convolution layers stacked together allow the system to build hierarchical feature representations: earlier layers detect basic edges, while deeper layers capture advanced features such as facial geometry and texture inconsistencies. ReLU introduces non-linearity, enabling the network to learn complex patterns beyond simple linear transformations. Multiple convolution layers stacked together allow the system to build hierarchical feature representations: earlier layers detect basic edges, while deeper layers capture advanced features such as facial geometry and texture inconsistencies.

$$Y(i,j) = \max_{(m,n) \subset R} X(i+m, j+n)$$

Pooling reduces dimensionality while retaining essential features, improving computational efficiency. By learning multi-level representations, convolutional layers enable VGG19 to detect subtle visual cues associated with deepfake images, making them indispensable in robust image manipulation detection systems.

b. Fully Connected (FC) Layers

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

Were,

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:

ISSN: 2349-6002

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

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

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

Were,

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

SoftMax ensures that all probabilities sum to 1, allowing the model to make a clear decision. Fully connected layers integrate all learned visual patterns such as facial edges, texture inconsistencies, and unnatural blending artifacts into a single classification outcome. In deepfake detection, this step is crucial because even small inconsistencies detected by convolutional layers influence the final probability score. The output is then forwarded to the blockchain module for secure, tamper-proof recording, ensuring transparent verification of detection results.

c. Recurrent Layer (RNN Layer)

The Recurrent Layer is the fundamental component of a standard Recurrent Neural Network (RNN). Unlike feedforward 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:

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

Were,

- 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 Layer (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 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 Networks (CNNs) and Recurrent Neural Networks (RNNs)

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

ISSN: 2349-6002

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:

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

Were,

- 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:

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

Were,

xt = CNN feature vector for frame t,

ht−1 = hidden state from the previous frame.

By combining frame-level CNN features with sequence-level RNN modeling, the system learns both the spatial authenticity of each frame and the temporal realism of the video as a whole. This hybrid CNN–RNN structure is especially effective because deepfake videos typically fail to maintain natural motion continuity, which the recurrent layer can easily detect.

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:

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

Where:

- TP (True Positives): Deepfakes correctly identified as fake.
- TN (True Negatives): Genuine content correctly identified as real.
- FP (False Positives): Real content incorrectly classified as fake.
- FN (False Negatives): Fake content incorrectly classified as real.

This formula evaluates the model's overall correctness across all prediction categories. For deepfake detection, maximizing accuracy means the system is consistently reliable in detecting manipulation while minimizing misclassification. However, accuracy alone may not fully reflect performance when datasets are imbalanced such as when real samples vastly outnumber fake ones. Even so, accuracy remains a crucial metric because it provides a straightforward assessment of how effectively the model performs in real-world scenarios, where both false positives and false negatives can have significant consequences.

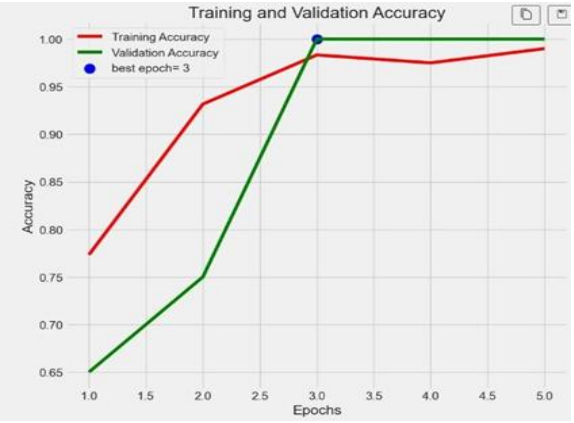

Fig. 4.1 Accuracy grapy for audio

ISSN: 2349-6002

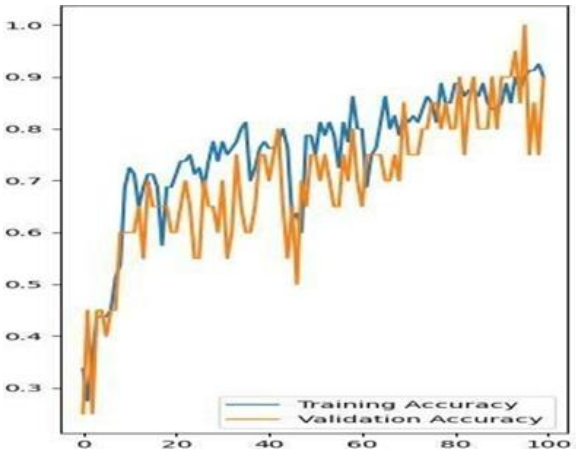

Fig. 4.2 Accuracy grapy for Video

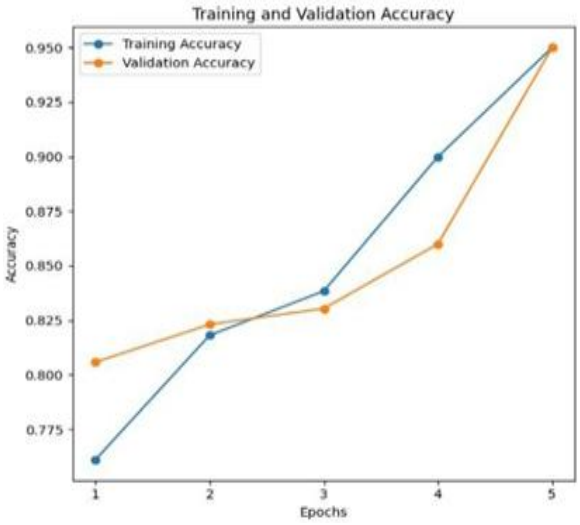

Fig. 4.3 Accuracy grapy for Image

The graph illustrating the training accuracy of image, audio, and video deepfake detection models highlights how each modality improves in performance as training progresses. The image model, based on VGG19-CNN, typically shows rapid accuracy growth in the early epochs because spatial features such as edges, textures, and facial details are easier for the network to learn. As training continues, the curve gradually stabilizes, indicating that the model has effectively learned to distinguish between real and manipulated visual patterns. The audio model, powered by LSTM and RNN, demonstrates a more gradual increase in accuracy because temporal patterns such as pitch, tone, and rhythm require more sequential learning. Its accuracy curve rises steadily as the model learns long-term dependencies and becomes better at detecting subtle synthesis artifacts in speech. The video model, combining CNN and RNN, generally requires more training time because it must learn both spatial and temporal features. Its accuracy curve starts lower but increases consistently as the model captures motion inconsistencies, lip-sync errors, and frame transitions.

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:

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

where:

- y = actual label (1 for fake, 0 for real),
- p = predicted probability that the sample is fake,
- 1−p = predicted probability that the sample is real.

This formula penalizes the model heavily when it assigns low probability to the correct class. For example, if the input is a deepfake (y=1) but the model predicts a very low probability ppp, the loss becomes large due to the log(p) term. Conversely, if the model predicts a probability close to 1 for the correct class, the loss becomes small. During training, the model updates its weights by minimizing this loss using optimization algorithms such as Adam or SGD. In multimodal systems, each branch image, audio, and video calculates its own loss, and these losses may be combined to improve joint learning performance. Monitoring the loss curve helps in understanding model convergence, detecting overfitting, and ensuring proper learning. A steadily decreasing loss curve indicates successful training, while fluctuating or increasing loss signals the need to adjust hyperparameters, learning rates, or model architecture.

ISSN: 2349-6002

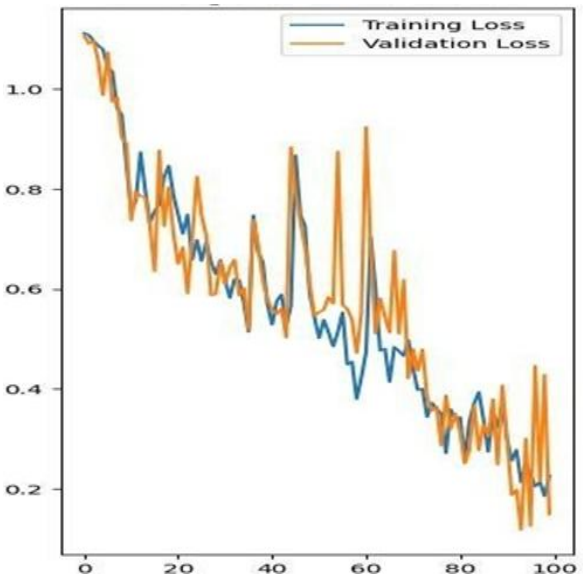

Fig. 4.4 loss grapy for Video

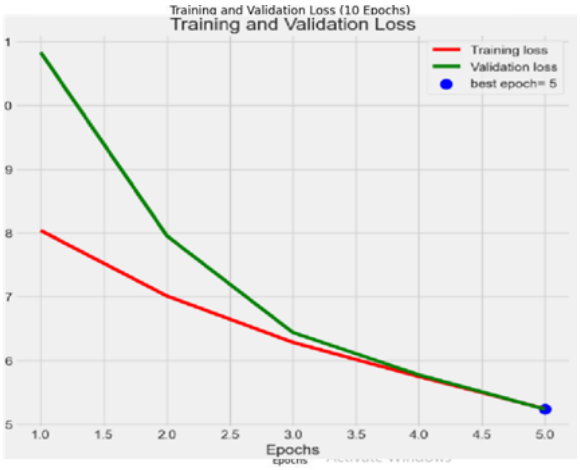

Fig. 4.5 loss grapy for audio

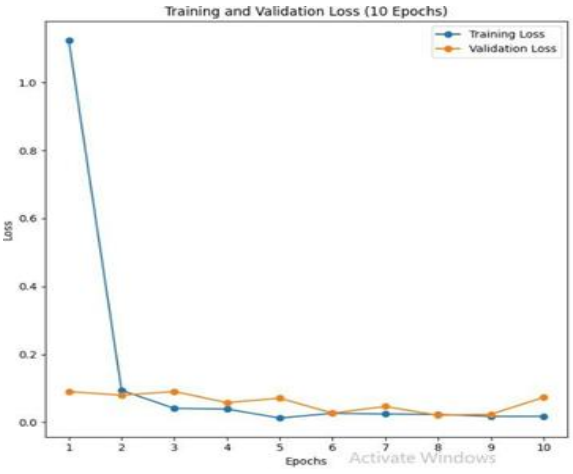

Fig. 4.5 loss grapy for Image

The graph representing the training loss for the image, audio, and video deepfake detection models shows how each modality reduces error as training progresses. The image model typically exhibits a sharp decline in loss during early epochs because CNNs quickly learn spatial features such as edges and textures. The audio model, built with LSTM and RNN layers, shows a more gradual loss reduction since learning temporal dependencies like pitch and rhythm requires more sequential processing. The video model, which combines CNN and RNN, generally displays the slowest initial loss decrease because it must learn both spatial and temporal patterns. Over time, all three loss curves stabilize at low values, indicating that each model has effectively minimized prediction errors and achieved convergence.

h. Precision:

Precision is an essential performance metric in deepfake detection because it measures how accurately the system identifies fake content without mistakenly labeling real content as fake. It reflects the model's ability to avoid false positives, which is extremely important in real-world scenarios where wrongly accusing genuine content of being fake can cause reputational harm or spread misinformation. Precision evaluates the proportion of correctly predicted deepfake samples out of all samples the model classified as fake. The mathematical formula for precision is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where:

- TP (True Positives) = deepfake samples correctly identified as fake,
- FP (False Positives) = real samples incorrectly identified as fake.

A high precision value means the system is reliable when it flags content as manipulated, ensuring that most predictions are correct. In multimodal deepfake detection, precision helps assess the performance of each model: CNNs tend to show strong precision in image detection due to clear spatial features, while LSTM/RNN audio precision improves as the model learns long-term voice patterns. Video precision typically increases when CNN and RNN layers are combined, reducing false alarms by considering both spatial and temporal cues.

i. Recall:

Recall is an important evaluation metric in deepfake detection because it measures the model's ability to correctly identify all fake content present in a dataset. It focuses on minimizing false negatives, which occur when deepfake images, audio, or videos are mistakenly classified as real. Missing a deepfake can be more harmful than a false alarm, especially in security, media verification, or forensic applications. Recall shows how effectively the system captures every manipulated sample. The formula for recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

- TP (True Positives) = deepfakes correctly identified as fake,
- FN (False Negatives) = deepfakes incorrectly labeled as real.

A high recall value means the model successfully detects most fake content, making it useful for applications where no deepfake should go unnoticed. In multimodal systems, recall helps measure how well image, audio, and video models identify all types of manipulations.

j. F1 Score

F1 Score is a crucial evaluation metric in deepfake detection because it provides a balanced measure of a model's performance by combining both precision and recall into a single value. While precision focuses on how many predicted deepfakes are actually fake, and recall measures how many fake samples the model successfully detects, the F1 score ensures neither metric is overlooked. This makes it especially useful when dealing with imbalanced datasets, which are common in deepfake detection. The F1 score is the harmonic mean of precision and recall, ensuring that a high value is achieved only when both metrics are strong. The formula for the F1 score is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high F1 score indicates that the model is both accurate in predicting fake content and effective at identifying all manipulated samples. This makes it a reliable overall performance indicator for image, audio, and video deepfake detection systems.

V. 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. The system securely maintains detection results and evidence within a controlled database environment, ensuring consistency, reliability, and proper record management. Together, these components create a robust, trustworthy, and scalable deepfake detection framework capable of safeguarding information integrity on social media platforms, restoring user confidence, and strengthening digital content authenticity in an increasingly AI-driven world. 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.

REFERENCE

[1] V. V. V. N. S. Vamsi, S. S. Shet, and S. S. M. Reddy, "Deepfake detection in digital media forensics," 2022.

[2] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," 2024.

[3] E. Nas and R. de Kleijn, "Conspiracy thinking and social media use are associated with ability to detect deepfakes," 2024.

[4] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," 2025.

[5] M. Hameleers, T. G. L. A. van der Meer, and T. Dobber, "Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes," 2024.

[6] M. Anil, K. Shiva, Y. Tulasi, G. Jayasri, B. Prudhvi Teja, and E. K. M. Sai, "Fake image

ISSN: 2349-6002

document detection via a deep discriminate model," International Research Journal of Engineering and Technology (IRJET), 2024.

[7] "Review on detection of GAN generated fake images over social networks," International Research Journal of Engineering and Technology (IRJET), 2024.

[8] A. Singh and J. Singh, "Image forgery detection using deep neural network," 2021.

[9] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A literature review," 2024.

[10] S. Migliorini, M. Gambini, and A. Belussi, "A blockchain-based platform for ensuring provenance and traceability of donations for cultural heritage," 2025.