Deepfake Detection: Deep Learning Based System for Identifying Synthetic Media

M.R.K Raju [1], Ch.Harika[2], Ch.S.D.Prem Kumar[3], B.M.L.Sudha[4], G.Iswarya[5]
[1]*Assistant professor, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-195132-459

Abstract: The rapid advancement of Artificial Intelligence (AI) and Deep Learning (DL) has significantly transformed digital media generation. Among these developments, deepfake technology has gained considerable attention due to its ability to create highly realistic manipulated images and videos. Deepfakes are generated using deep neural networks that can replace faces, modify expressions, or alter speech patterns with convincing accuracy. Although this technology has useful applications in entertainment, filmmaking, and virtual environments, its misuse for misinformation, identity fraud, cyber harassment, and political manipulation has raised serious ethical and security concerns. This has created an urgent need for reliable deepfake detection systems.

This research proposes a deep learning-based framework for identifying manipulated media using a hybrid CNN-LSTM architecture. The model employs a pretrained MobileNetV2 Convolutional Neural Network (CNN) to extract spatial features from image frames, while a Long Short-Term Memory (LSTM) network captures temporal dependencies across video sequences. The system processes both images and videos by extracting twenty representative frames, resizing them to 224*224 pixel, and applying normalization before classification. The dataset used for training and evaluation includes balanced samples from the Celeb-DF, and Deepfake Detection Challenge (DFDC) datasets. The model was implemented using TensorFlow and trained with GPU support in Google Collab. Experimental results indicate an approximate accuracy of 80%, demonstrating effective discrimination between real and manipulated content. A Gradio-based interface was also developed to enable practical media verification.

Keywords: Deepfake Detection, CNN-LSTM, Synthetic Media, Temporal Analysis, Digital Forensics

I. INTRODUCTION

The rapid advancement of AI has significantly changed the way digital content is created and shared. Among these developments, deepfake technology has emerged as one of the most impactful and controversial innovations. Deepfakes are artificially generated images or videos created using deep learning models that can realistically alter facial identity, expressions, lip movements and even voice. With the advancement of generative models, especially Generative Adversarial Networks (GANs), it has become possible to produce highly convincing fake video that are difficult to distinguish from genuine recordings.

Although deepfake technology has useful applications in filmmaking, virtual characters, gaming, and other entertainment industries, its misuse has created serious social and security challenges. Manipulated videos have been used to spread false information, fabricate political statements, perform financial scams, and harm individual reputations. The availability of open-source software and high-performance computing resources has made this technology accessible to a wider population, reducing the technical barrier required to create synthetic media. Consequently, verifying the authenticity of digital content has become increasingly important.

Earlier detection approaches relied on manual inspection or simple visual artifact analysis, which are no longer sufficient due to improvements in deepfake quality. Modern manipulated video exhibit strong visual consistency with minimal visible distortions. Therefore, automated detection systems are necessary to identify subtle inconsistencies that may not be noticeable to the human eye. DL techniques have demonstrated strong capability in extracting complex patterns from images and videos. However, several existing methods primarily focus on analyzing individual frames and often neglect temporal relationships between consecutive frames.

To address this limitation, this research proposes a hybrid deep learning framework that combines spatial

and temporal analysis. A pretrained MobileNetV2 convolutional neural network is used for extracting spatial features, while a Long Short Term Memory network models the sequential dependencies across frames. The objective is to develop an efficient and practically deployable deepfake detection deepfake detection system capable of analyzing both images and videos with reliable accuracy.

II. LITERATURE SURVEY

Deepfake detection has become an important research area in computer and digital forensics. As generative models improved in realism and quality, detection techniques evolved to address more sophisticated manipulations. Early research mainly focused on identifying visible artifacts introduced during face-swapping processes. Researchers examined inconsistencies such as abnormal lighting, unnatural facial boundaries, distorted skin textures, and mismatched color tones . Although these handcrafted feature-based methods were effective for detecting low-quality deepfakes, they were not robust against advanced generative techniques.

With the development of CNNs, researchers began applying deep learning models for automated feature extraction and classification. Architectures such as VGGNet, ResNet, and Inception were used to detect manipulated images by learning discriminative spatial patterns directly from data. These CNN-based approaches achieved better performance than traditional manual methods. However, most of these models processed frames independently, limiting their ability to detect temporal inconsistencies in videos.

As deepfake videos became increasingly realistic, researchers emphasized the need for temporal modeling. Even when individual frames appear authentic, subtle irregularities may occur across consecutive frames, such as unnatural blinking patterns, inconsistent facial movements, flickering artifacts, or abnormal head pose transitions. To address this issue, recurrent neural networks, particularly LSTM models, were introduced to capture sequential dependencies. Hybrid CNN-LSTM architectures demonstrated improved performance by combining spatial feature extraction with temporal sequence modeling.

Benchmark datasets such as Celeb-DF and the Deepfake Detection Challenge (DFDC) significantly supported research progress. These datasets include manipulated videos created using different techniques and compression settings, enabling comprehensive evaluation of detection models. Studies indicates that properly optimized lightweight architectures can achieve competitive accuracy.
Recent research also highlights the importance of computational efficiency and deploy ability.

While deep architectures like EfficientNet and transformer –based models offer high accuracy, they require substantial resources. Lightweight models such as MobileNetV2provide a balanced trade-off between accuracy and efficiency by using depth-wise separable convolutions and inverted residual blocks.

Despite considerable progress, achieving strong generalization across diverse manipulation methods and compression artifacts remains challenging. Therefore, integrating efficient convolutional networks with temporal modeling remains a promising approach. Based on these insights, the proposed work adopts a pretrained MobileNetV2 backbone combined with an LSTM network to enhanced detection reliability while maintaining computational feasibility.

III. SYSTEM ANALYSIS

System analysis plays a crucial role in understanding the functional requirements of a detection framework and identifying the limitations of currently available solution. With the rapid improvement of generative models, deepfake content has become increasingly realistic, making detection conventional verification methods insufficient. Therefore, it is important to examine existing systems before proposing an improved and practical solution.

a.Existing System
Most existing deepfake detection approaches rely primarily on frame analysis using CNN. In such systems, individual frames are extracted from videos and classified as real or manipulated based on spatial features. While this strategy performs reasonably well for low-quality deepfakes, it becomes less effective when handling high-resolution synthetic content generated using advanced GAN-based techniques.

A major limitation of these systems is the lack of temporal analysis. Since modern deepfake videos often strong visual consistency in individual frames, frame-wise classification alone cannot reliably detect manipulation. Subtle motion inconsistencies, unnatural facial transitions, and irregular temporal patterns frequently remain unnoticed in purely spatial models.
Another concern with many state-of-the-art approaches is computational complexity. Deep architectures with a large number of parameters demand significant processing power and memory resources. This restricts their deployment in real time or resource-constrained environments. Additionally, certain traditional detection methods depends on handcrafted features or biological cues such as eye-blinking frequency. As generative techniques continue to advance, these handcrafted methods become less reliable.

Therefore, existing systems face several key challenges:

- Inability to capture temporal dependencies.
- Limited generalization across diverse manipulation techniques
- High computational overhead
- Reduced suitable for real-world deployment

These limitations emphasize the need for a more balanced and efficient detection framework.

b. Proposed System

To address the shortcomings of existing approaches, this research proposes a hybrid deep learning-based detection system that integrates both spatial and temporal analysis within a lightweight architecture.
The proposed framework utilize MobileNetV2 as a pretrained convolutional backbone to extract spatial features from input media. MobileNetV2 was selected due to its computational efficiency and strong feature representation capability achieved through depth-wise separable convolutions.
To incorporate temporal modellings, features extracted from twenty representative frames are passed to a LSTM network. The LSTM layer analyzes sequential dependencies and identifies inconsistencies that may not be detectable when frames are processed independently.
The system is designed to process both images and videos. For image inputs, a sequence simulation strategy is applied to maintain architectural consistency. For video inputs, evenly spaced frame sampling ensures that the entire duration of the video is represented fairly
The proposed system aims to achieve:

- Enhanced detection accuracy through spatial-temporal modeling
- Reduced computational complexity using lightweight CNN
- Capability to analyze both images and videos
- Practical deployment through a user-friendly interface

By combining efficiency with performance, the proposed solution offers a scalable and reliable approach for deepfake detection

IV. SYSTEM DESIGN AND ARCHITECTURE

The proposed system follows a modular architecture to ensure clarity, scalability, and efficient data processing. The workflow begins with media input acquisition and proceeds through preprocessing, spatial feature extraction, temporal modeling, classification, and final output generation. Each module performs a defined task, allowing smooth data flow throughout the system.
The input model accepts either image or video files. For image inputs, the media is resized to 224 *224 pixels to match the required input dimensions of the convolutional backbone. Since the architecture expects sequential data, the image is replicated twenty times to simulate a temporal sequence. For video inputs, twenty evenly spaced frames are extracted using linear sampling strategy to ensure proper coverage of the entire video duration. This enables the model to analyze both spatial and temporal characteristics effectively.

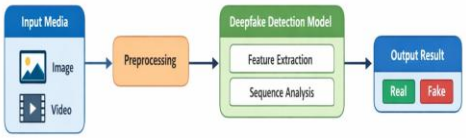

Fig. 1. Overall Architecture of the Proposed Deepfake Detection System

The preprocessing module normalizes pixel values using the transformation (pixel value / 127.5) – 1.0. This step ensures compatibility with MobileNetV2 standards and improves training stability.
The feature extraction stage utilizes MobileNetV2 pretrained on ImageNet. The network extracts high-level spatial features such as texture inconsistencies,

blending artifacts, and structural irregularities in facial regions. Due to its lightweight design and depth-wise separable convolutions, MobileNetV2 reduce computational cost while maintaining effective feature representation.

The temporal modeling module consists of a LSTM layer. This layer processes the sequence of feature vector obtained from twenty frames and captures temporal dependencies. By analyzing motion continuity and frame transitions, the LSTM identifies inconsistencies that may indicate manipulation.

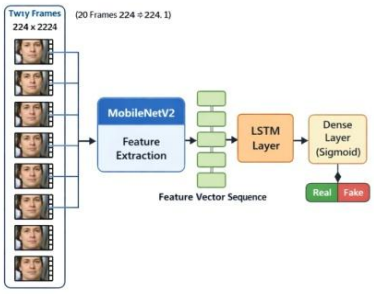

Fig.2.Proposed CNN-LSTM Deepfake Detection Architecture

The classification module includes a dense layer with sigmoid activation, producing a probability score between 0 and 1. A predefined threshold is applied to classify the input as real or fake, and a confidence percentage is calculated to improve interpretability.

Finally, the deployment module integrates the trained model with a Gradio-based user interface. This interface enables real-time media verification, allowing users to upload content and receive authenticity predictions efficiently. The modular structure supports practical usability and future scalability.

V.IMPLEMENTATION

The proposed deepfake detection system was implemented using Python and the TensorFlow framework. The development process included dataset preparation, preprocessing, model construction, training, evaluation, and deployment. All experiments were performed in Google Colab with GPU acceleration to improve computational efficiency.

The dataset consists of samples from the Celeb-DF and DeepFake Detection Challenge (DFDC) datasets. A balanced subset containing 250 real and 250 manipulated videos was selected to ensure unbiased training. The data was divided into training and validation sets to evaluate model generalization, and shuffling was applied to reduce ordering bias.

For video inputs, twenty frames were extracted using a linear sampling strategy to ensure proper coverage of the entire video duration. Frame extraction was performed using OpenCV, and each frame was resized to 224 × 224 pixels to match the input requirements of MobileNetV2.

During preprocessing, pixel values were normalized using the transformation:

Normalized Pixel = (Pixel Value / 127.5) – 1.0

This normalization aligns the input distribution with MobileNetV2 standards and supports stable training. The model architecture uses a pretrained MobileNetV2 network as the spatial feature extractor. Transfer learning was applied by initializing the network with ImageNet weights. The spatial features extracted from twenty frames were arranged sequentially and passed to a Long Short-Term Memory (LSTM) layer to capture temporal dependencies and motion inconsistencies across frames.

The final classification layer consists of a dense neuron with sigmoid activation to perform binary prediction. The model was compiled using Binary Cross-Entropy loss and optimized with the Adam optimizer. Training was conducted for multiple epochs while monitoring validation accuracy to reduce overfitting.

After training, the model was saved and integrated into a Gradio-based web interface. The deployment module allows users to upload image or video files and receive authenticity predictions in real time. For video inputs, a format conversion step ensures playback compatibility. Overall, the implementation effectively combines spatial and temporal modeling while maintaining computational efficiency suitable for practical applications.

VI.RESULTS AND DISCUSSION

The performance of the proposed CNN–LSTM model was evaluated using the validation dataset. The evaluation mainly focuses on metrics such as accuracy, loss trends, and overall model performance. The experimental results indicate that the model achieved an approximate classification accuracy of around 80% on the validation dataset, which demonstrates the capability of the system to distinguish between real and

manipulated media samples.

During the training process, both accuracy and loss values were monitored across multiple epochs to understand the learning behavior of the model. The combined accuracy and loss graph shown in Fig. 3&4 illustrates the variation of training accuracy and loss during the training phase.

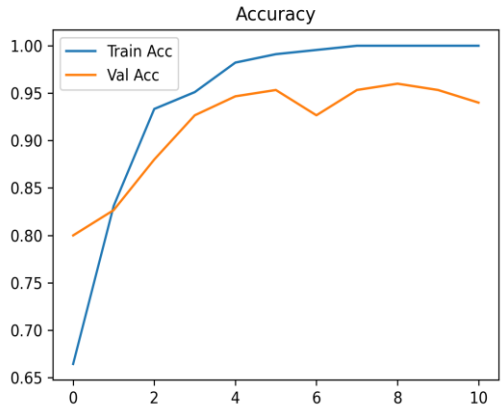

Fig:3 Training Accuracy Graph

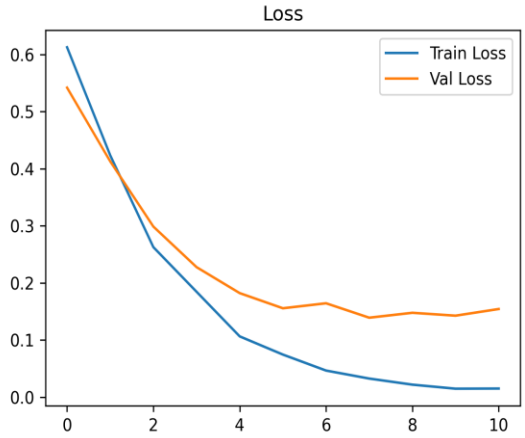

Fig:4 Training Loss Graph

The graph shows that the accuracy gradually increases while the loss value decreases as the number of epochs increases. This indicates that the model is effectively learning meaningful patterns from the dataset and improving its prediction capability.

The overall performance of the proposed system is illustrated in Fig. 5, which represents the performance graph of the trained model. The graph highlights the system's ability to correctly classify a majority of the input samples as either real or fake. Although minor misclassifications may occur in certain cases due to subtle manipulations or compression artifacts, the model still maintains satisfactory detection performance.

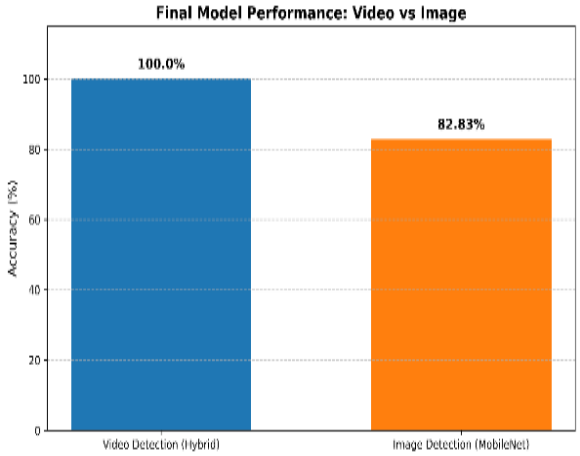

Fig:5 Performance Graph of the Proposed Model

From a computational perspective, the use of MobileNetV2 ensures a lightweight architecture with reduced computational overhead and faster inference time. This makes the proposed system suitable for practical deployment scenarios where efficient deepfake detection is required. Overall, the results confirm that the hybrid CNN–LSTM approach effectively combines spatial feature extraction and temporal analysis to detect manipulated media.

VII.CONCLUSION

This research presented a deep learning-based framework for detecting deepfake images and videos using a hybrid CNN–LSTM architecture that combines spatial and temporal analysis. The system integrates MobileNetV2 for efficient spatial feature extraction and a Long Short-Term Memory network to model sequential dependencies across video frames. By examining both frame-level details and motion continuity, the proposed approach improves detection reliability compared to traditional frame-based methods. The model was trained and evaluated using balanced samples from the Celeb-DF and DeepFake Detection Challenge (DFDC) datasets, achieving an approximate accuracy of 80%. The use of transfer learning reduced training time and enhanced generalization performance.

In addition, the lightweight design of MobileNetV2 ensures lower computational complexity, making the system suitable for practical deployment. The integration of a Gradio-based interface demonstrates real-world applicability by allowing users to upload media and receive authenticity predictions along with

confidence scores. Although certain challenges remain, particularly in detecting highly sophisticated manipulations and compressed content, the overall framework provides a balanced and scalable solution. Future work may focus on expanding the dataset, incorporating attention mechanisms, and exploring advanced architectures to further strengthen detection robustness in evolving digital environments.

REFERENCES

[1] I. Goodfellow et al., "Generative Adversarial Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] A. Howard et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] H. Li, B. Li, and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," *IEEE CVPR Workshops*, 2019.

[5] Y. Li et al., "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," *IEEE CVPR*, 2020.

[6] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," arXiv:2006.07397, 2020.

[7] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," *IEEE AVSS*, 2018.

[8] T. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," *IEEE WIFS*, 2018.

[9] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," *IEEE ICCV*, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE CVPR*, 2016.

[11] C. Szegedy et al., "Going Deeper with Convolutions," *IEEE CVPR*, 2015.

[12] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *ICML*, 2019.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NeurIPS*, 2012.

[14] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *IEEE CVPR*, 2017.

[15] R. Tolosana et al., "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.