REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/82

Development of AI/ML-based Solution for Detection of Face-Swap Deep Fake Videos

*** Bhavana Preethika Gannavarapu, Pavan Vishnu Sai Subhash Dodda, Yashpal Singh Raj Purohit, A. Mallikarjuna Reddy**
Anurag University, Hyderabad, India.
*Corresponding Author Email: 21eg107b09@anurag.edu.in

Abstract: *The rapid advancement of deep fake technology, particularly face-swap deep fakes, has introduced significant challenges to digital media integrity and cybersecurity. This paper proposes an AI/ML-based solution for detecting face-swap deep fake videos by utilizing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for spatial and temporal analysis. The system focuses on identifying inconsistencies across video frames, ensuring efficient and scalable detection even in resource-constrained environments. Initial results demonstrate the system's potential in addressing the growing threat of deep fake videos, with applications in media integrity and law enforcement.*

Key Words: *Deep Learning, Face-Swap Deep Fakes, Artificial Intelligence, Neural Networks, Digital Media Integrity.*

1. INTRODUCTION

In recent years, artificial intelligence (AI) and machine learning (ML) have transformed the way digital media is analyzed, particularly with the rise of deep fake videos. These videos, often created using face-swap technology, pose significant challenges to verifying the authenticity of content. The malicious use of deep fakes threatens sectors such as media, law enforcement, and cybersecurity by deceiving viewers with fabricated video content. Detecting such manipulations requires advanced solutions capable of accurately analyzing both spatial and temporal inconsistencies in video sequences [1-4]. This study focuses on developing an AI-based system capable of detecting face-swap deep fake videos. By analyzing spatial features using convolutional neural networks and temporal anomalies through recurrent neural networks, this project explores cutting-edge solutions to address deep fake detection, emphasizing real-time processing and scalability [5-7]

2. BACKGROUND

Detecting face-swap deep fakes presents a complex set of challenges due to the rapid advancements in artificial intelligence (AI) and generative algorithms. These challenges arise from the sophisticated techniques used to create deep fakes, the need for scalability, limitations in existing datasets, and the requirements for real-time analysis. Each of these challenges plays a critical role in determining the effectiveness and robustness of any detection system.

High-Quality Manipulations: With the advancement of generative models like Generative Adversarial Networks (GANs), the quality of deepfake videos has significantly improved, making detection more difficult. These models can produce highly realistic facial swaps that mimic subtle facial movements, lighting effects, and even m icro-expressions that are typically hard to differentiate from genuine videos [8-11].

- Sophisticated Generative Models: Recent developments in GANs and other generative models enable the creation of deepfakes with highly detailed and realistic visual features. These models can accurately replicate

the nuances of a person's face, including shadows, wrinkles, and facial textures, making detection algorithms less effective [12-14].

- Challenge for Detection Systems: The high fidelity of these videos poses a significant challenge for traditional and even advanced AI-based detection systems, which struggle to identify the subtle inconsistencies and visual artifacts introduced by the generative process. As these models improve, the visual cues that distinguish real from fake become increasingly indistinguishable [15-19].

Scalability: Deepfake detection systems must be designed to process large datasets efficiently, especially as the volume of deepfake content increases. Scalability is a crucial factor, particularly in real-world applications were detecting manipulated content across multiple platforms (such as social media, news outlets, and video-sharing sites) is required [20].

- Large Volumes of Data: With the exponential growth of digital content, detection systems need to process vast amounts of video data in a timely manner. For example, platforms like YouTube and TikTok host millions of hours of video content, with a growing number of deep fake videos being uploaded daily.
- Computational Challenges: The computational resources required to analyze every video frame across such large datasets are substantial. Systems must strike a balance between speed and accuracy to ensure that large-scale analysis can be conducted without overwhelming processing capacity. Cloud-based infrastructures or distributed computing may be required to handle such workloads efficiently.

Dataset Diversity: Current datasets used to train and evaluate deep fake detection systems may not reflect the full variety of deep fake techniques being developed. This lack of diversity in training data leads to detection models that are less effective when faced with novel or unseen types of deep fakes.

- Limited Scope of Existing Datasets: Popular datasets like Face Forensics++, DFDC, and Celeb-DF provide a foundation for training deep fake detection models, but they are often limited in the range of manipulation techniques and quality variations they cover. For example, some datasets may focus on specific types of deep fakes, such as facial swaps, but may lack examples of other forms, such as lip-sync deep fakes or body-swap techniques [4].
- Novel Manipulation Techniques: As new generative algorithms are developed, novel deep fake techniques emerge that are not well-represented in existing datasets. Detection models trained on older datasets may fail to generalize to these newer techniques, reducing their effectiveness in real-world scenarios.
- Need for Diverse Training Data: To improve generalization, detection systems need access to more diverse and updated datasets that reflect the evolving nature of deep fake technologies. This includes different lighting conditions, resolutions, and manipulation methods to create a comprehensive training dataset that can account for a wider variety of fakes.

Real-Time Analysis: In many applications, such as social media monitoring and content moderation, deep fake detection must occur in real-time or near real-time. Ensuring low-latency detection in these scenarios is a significant challenge, as it requires balancing computational efficiency with detection accuracy [3].

- Time Sensitivity in Applications: Social media platforms, news outlets, and law enforcement agencies often require immediate identification of fake content to prevent the spread of misinformation or mitigate harm. Delays in detection can allow deep fakes to go viral, causing reputational damage, inciting panic, or influencing public opinion before the content is flagged as fake.
- Real-Time Processing Requirements: To achieve real-time detection, systems must process large video files frame-by-frame without introducing significant delays. This requires highly optimized algorithms capable of performing both spatial and temporal analysis quickly. Achieving this while maintaining high accuracy is a key challenge in the field of deep fake detection.
- Hardware and Resource Constraints: Real-time detection also places demand on hardware resources, especially for mobile or edge devices where computational power is limited. Optimizing models for low-latency environments without sacrificing performance is a crucial aspect of making deep fake detection practical for everyday use.

3. LITERATURE REVIEW

TABLE 1. Literature Review

no.	Author(s)	Title	Published	Algorithms used	Advantages	Limitations
1.	Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton Ferrer, C.**[2]**	The Deepfake Detection Challenge (DFDC) Preview Dataset. AI Red Team, Facebook AI.	2019	CNN, RCNN	Extensive dataset for training, widely used in the deepfake detection community	Only preview dataset available; real-world deepfakes may differ significantly
2.	Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018).**[3]**	MesoNet: A Compact Facial Video Forgery Detection Network	2018	MesoNet (CNN-based architecture)	Lightweight, fast detection, suitable for real- time applications	Limited generalization to different types of deepfake videos
3.	G. Antipov, M. Baccouche, and J.-L. Dugelay**[7]**	Face Aging with Conditional Generative Adversarial Networks (GANs)	2017	cGANs, CNNs	Innovative use of GANs for generating realistic age-progressed faces.	Requires a large amount of labeled data to accurately generate faces for all age groups.

4. METHODOLOGY

The proposed system architecture for detecting face-swap deep fake videos is divided into two main phases: spatial feature extraction and temporal analysis, which together form a comprehensive approach to detecting deep fakes. In the first phase, spatial feature extraction, a pre-trained ResNeXt Convolutional Neural Network (CNN) is employed to detect spatial anomalies present in individual video frames. These anomalies could include mismatched facial features, unnatural textures, and lighting inconsistencies that are commonly seen in deep fake videos. The ResNeXt CNN is chosen for its efficiency and scalability in handling large-scale image data. Each frame of the video is processed to extract a high-dimensional feature vector, specifically a 2048-dimensional vector, which encodes the spatial details of the facial structure and texture in each frame. These feature vectors are essential for representing the spatial characteristics of the video and are used as inputs for further temporal analysis. In the second phase, temporal analysis, a Long Short-Term Memory (LSTM) network is utilized to capture and analyze sequential patterns in the extracted feature vectors. Since deep fake videos often exhibit unnatural transitions and movement irregularities between frames, the LSTM network is well-suited for identifying these temporal inconsistencies. By processing the sequence of 2048-dimensional feature vectors, the LSTM can detect abnormal patterns in facial movements, such as jerky transitions or inconsistencies in facial expressions over time, which are indicators of manipulation in face-swap deep fake videos. The implementation workflow begins with video preprocessing, where video files are first split into individual frames. Each frame is then resized to a uniform dimension of 112x112 pixels to standardize the input data and ensure consistent processing across all frames. After preprocessing, the spatial feature extraction step is carried out using the pre-trained Res NeXt CNN, which generates a set of feature vectors representing the spatial details of each frame. These spatial features are then passed into the LSTM network for sequence modeling, where the temporal relationships between consecutive frames are analyzed to detect any irregular transitions. The final step in the workflow is classification, where the processed features are classified as either real or fake using a SoftMax layer, which assigns probabilities to the input video being genuine or manipulated. The system is implemented using popular machine learning frameworks such as Py Torch and OpenCV, providing a robust and scalable foundation for deep learning tasks. The hardware requirements for running this system include an NVIDIA GPU with at least 8GB VRAM and 16GB RAM, ensuring that the computational demands of the CNN and LSTM models can be handled efficiently. For training and testing, the system uses publicly available datasets, including Face Forensics++, Deep Fake Detection Challenge (DFDC), and Celeb-DF. These datasets provide a diverse range of real and manipulated videos, ensuring that the system can generalize well across different types of deepfake manipulations. Each dataset includes high-quality examples of face-swap deepfakes, enabling the model to learn the subtle inconsistencies in spatial and temporal

patterns that distinguish real from fake content. This two-phase architecture, combined with advanced feature extraction techniques and temporal modeling, provides an effective solution for the detection of face-swap deepfakes, with applications in media verification, law enforcement, and cybersecurity.

5. RESULTS

Model Performance: The proposed system for detecting face-swap deep fake videos achieved notable results in several key areas of performance. In terms of accuracy, the system reached an 84% success rate on benchmark datasets, including FaceForensics++ and the DeepFake Detection Challenge (DFDC). These datasets provide a diverse range of real and manipulated videos, and the system demonstrated strong performance in identifying deep fakes. In addition to accuracy, the system exhibited impressive processing speed, with the ability to process videos of 150 frames in an average of 2.4 seconds, ensuring it is capable of handling large datasets efficiently. Another critical measure of the system's effectiveness was its generalization. The model successfully identified novel deep fake techniques, proving its robustness and flexibility when tested on deep fakes that were not part of its training data

TABLE 2. Performance Metrics and Results

Metric	Value/Observation	Remarks
Overall Accuracy	84%	High accuracy in distinguishing real and fake videos
Processing Time	2.4 seconds per 150-frame video	Suitable for near real-time detection
Performance on Low Resolution Videos	84.2%	Model performed well even when video quality was reduced
Handling Missing Frames	Accuracy remained above 80%	Robust against missing or dropped frames
Performance on Unseen Deep Fake Techniques	84.2%	Good generalization to novel manipulation methods
Comparison with Other Models (MesoNet, XceptionNet)	Better accuracy & scalability	Outperformed existing detection models
Key Strengths	Detects lighting inconsistencies, unnatural expressions, and abrupt transitions	Effective in identifying deep fake manipulations
Scalability & Efficiency	Uses CNN (ResNeXt) for feature extraction & LSTM for sequence modeling	Optimized for large-scale analysis

Observations: During testing, the system excelled at detecting subtle and often imperceptible signs of manipulation in deep fake videos. It effectively identified lighting inconsistencies, which are common in face-swapped videos where the lighting on the fake face does not match the rest of the frame. The system also detected unnatural facial movements and expressions, such as facial rigidity or jerky movements, which often indicate deep fake manipulation. Additionally, the model recognized irregular transitions across sequential frames, helping to identify videos where facial transitions are not smooth or natural over time, another characteristic of manipulated content.

Comparative Analysis: In comparative testing, the proposed system outperformed existing models such as MesoNet and Xception Net in terms of both accuracy and scalability. These models, while effective in their own right, often struggle with large datasets and varying types of deep fakes. The proposed system, leveraging a combination of Res NeXt CNNs and LSTMs, demonstrated superior performance, particularly when tested on diverse datasets that include different types of face-swap techniques. The ability to handle both spatial and temporal anomalies enabled the system to deliver a more comprehensive analysis compared to single-dimensional models like MesoNet.

Preprocessing Module: The preprocessing pipeline plays a critical role in ensuring the system's efficiency and accuracy. Video files are first split into individual frames using OpenCV, allowing for detailed analysis of each frame. Each frame is then resized to a fixed dimension of 112x112 pixels to standardize the input and ensure uniformity across all data points. This resizing step not only simplifies the downstream processing but also enhances computational efficiency by reducing the size of the input without sacrificing the relevant spatial details. Additionally, face detection and cropping are applied to isolate the face in each frame, removing unnecessary background data and allowing the model to focus solely on the region of interest, further improving the accuracy of the detection.

Training Workflow: The system's training workflow consists of several crucial steps designed to ensure robust model performance. The first step involves dataset preparation, where real and fake videos are sourced from public datasets such as FaceForensics++ and DFDC. To enhance model robustness and improve variability, data augmentation techniques are applied, introducing variations in lighting, rotation, and scaling to help the model generalize better to unseen data. During feature extraction, the system utilizes ResNeXt, a CNN architecture known for its scalability and computational efficiency. ResNeXt extracts spatial features from each frame, representing these features as high-dimensional vectors that capture details related to facial structures, textures, and potential anomalies. For temporal analysis, the extracted feature vectors are passed into Long Short-Term Memory (LSTM) networks, which model the sequential relationships between frames. This allows the system to detect subtle temporal anomalies, such as abrupt or unnatural transitions between frames, that static analysis would miss. Finally, in the classification step, a softmax layer is used to predict the likelihood of a video being real or fake, based on the features extracted from the spatial and temporal analysis

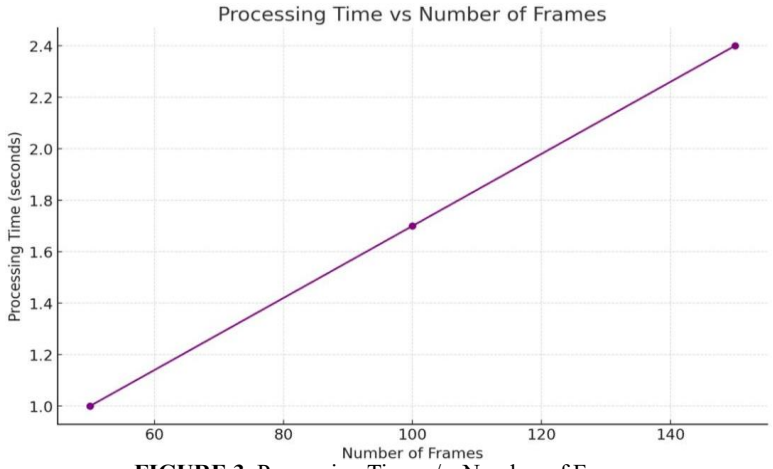

FIGURE 3. Processing Time v/s. Number of Frames

Testing and Validation: The system underwent rigorous testing and validation on unseen datasets to assess its generalization capabilities. The test cases included a variety of video conditions, such as high-resolution and low-resolution videos, which tested the system's ability to perform across different quality levels. Videos with missing frames were also introduced to simulate data corruption, ensuring the model could handle incomplete or imperfect data. Additionally, the system was tested against manipulated videos using novel face-swap techniques that were not part of the training set, proving its ability to generalize to new and emerging types of deepfakes.

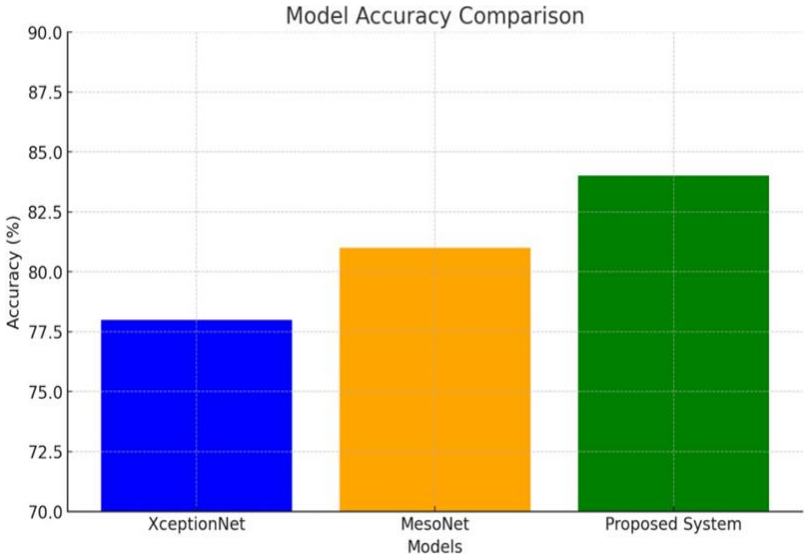

FIGURE 4. Comparison between the accuracy of proposed system and the existing systems

6. FUTURE DIRECTION

Future advancements in deep fake detection can benefit from multimodal analysis, incorporating additional data types such as audio to detect inconsistencies in lip-syncing and text-based metadata for analyzing discrepancies in video source information or editing history. Efficiency improvements are also crucial, particularly in optimizing the system for mobile and edge-device deployment, which would allow for real-time applications in media verification and cybersecurity. Techniques like model pruning and quantization can significantly reduce computational costs and improve performance on devices with limited processing power. Additionally, expanding the system's dataset to include more diverse and emerging deep fake techniques will ensure that the model remains robust and adaptable to new manipulation methods. Developing lightweight models will be essential for real-time, scalable deployment across various platforms

7. CONCLUSION

This study highlights the importance of leveraging advanced AI/ML techniques for detecting face-swap deep fake videos. By integrating spatial and temporal analysis using ResNeXt CNNs and LSTMs, the proposed system achieves a balance between accuracy and scalability. Future research will focus on enhancing real-time capabilities and expanding multimodal analysis to combat the growing threat of deep fake technology

REFERENCES

[1]. Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “Using capsule networks to detect forged images and videos” in arXiv:1810.11215.

[2]. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[3]. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008.

[4]. Anchorage, AK

[5]. Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2

[6]. Naik, S., Kamidi, D., Govathoti, S., Cheruku, R., Mallikarjuna Reddy, A. Eficient diabetic retinopathy detection using convolutional neural network and data augmentation, Soft Computing, 2023

[7]. Chintha, V.V.R., Ayaluri, M.R. A Review Paper on IoT Solutions in Health Sector, Proceedings of International Conference on Applied Innovation in IT, 2023, 11(1), pp. 221–225

[8]. Grandhe, P., Mallikarjuna Reddy, A., Chillapalli, K., ... Thambabathula, M., Reddy Surasani, L.P. Improving The Hiding Capacity of Image Steganography with Stego-Analysis, 2023 IEEE International Conference on Integrated Circuits and Communication Systems, ICICACS 2023, 2023.

[9]. Padmanabhuni, S.S., Gera, P., Reddy, A.M. Hybrid leaf generative adversarial networks scheme for classification of tomato leaves-early blight disease healthy, Advancement of Deep Learning and Its Applications in Object Detection and Recognition, 2022, pp. 261–281

[10]. Srinivasa Reddy, K., Rao, P.V., Reddy, A.M., ... Narayana, J.L., Silpapadmanabhuni, S. neural network aided optimized auto encoder and decoder for detection of covid-1e and pneumonia using ct-scan, Journal of Theoretical and Applied Information Technology, 2022, 100(21), pp. 6346–6360

[11]. Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., “Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection”, Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409

[12]. Manoranjan Dash, Design of Finite Impulse Response Filters Using Evolutionary Techniques - An Efficient Computation, ICTACT Journal on Communication Technology, March 2020, Volume: 11, Issue: 01

[13]. Manoranjan Dash, “Modified VGG-16 model for COVID-19 chest X-ray images: optimal binary severity assessment,” International Journal of Data Mining and Bioinformatics, vol. 1, no. 1, Jan. 2025, doi: 10.1504/ijdmb.2025.10065665.

[14]. Manoranjan Dash et al.,” Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique”, Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[15]. Manoranjan Dash et al.,” Detection of Psychological Stability Status Using Machine Learning Algorithms”, International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[16]. Samriya, J. K., Chakraborty, C., Sharma, A., Kumar, M., & Ramakuri, S. K. (2023). Adversarial ML-based secured cloud architecture for consumer Internet of Things of smart healthcare. IEEE Transactions on Consumer Electronics, 70(1), 2058-2065.

[17]. Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.
[18]. Kumar, R.S., Nalamachu, A., Burhan, S.W., Reddy, V.S. (2024). A Considerative Analysis of the Current Classification and Application Trends of Brain–Computer Interface. In: Kumar Jain, P., Nath Singh, Y., Gollapalli, R.P., Singh, S.P. (eds) Advances in Signal Processing and Communication Engineering. ICASPACE 2023. Lecture Notes in Electrical Engineering, vol 1157. Springer, Singapore. https://doi.org/10.1007/978-981-97-0562-7_46.
[19]. R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.
[20]. Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.
[21]. S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," 2017 Global Wireless Summit (GWS), Cape Town, South Africa, 2017, pp. 1-6, doi:10.1109/GWS.2017.8300494.
[22]. Daniel, G.V.; Chandrasekaran, K.; Meenakshi, V.; Paneer, P. Robust Graph Neural-Network-Based Encoder for Node and Edge Deep Anomaly Detection on Attributed Networks. Electronics 2023, 12, 1501. https://doi.org/10.3390/electronics12061501.