International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Use of LSTM Machine Learning Integrated with Explainable AI (x-AI) in Face Deep-fake Detection

Abhinava Karthic, Harshit Kupati, M. Chaithanya, Pradnya Kanamitte, Dr. Rajgopal K.

Department of Computer Science and Engineering, S.E.A College of Engineering & Technology, Bengaluru, VTU

Abstract

This paper presents a novel approach for detecting face deepfakes using Long Short-Term Memory (LSTM) networks integrated with Explainable AI (x-AI) techniques. With the rise of synthetic media generated by AI, particularly deepfake videos and images, there is an urgent need for robust detection systems. Our work utilizes the temporal capabilities of LSTM networks to learn temporal facial patterns, while incorporating explainable methods such as SHAP and LIME to offer transparency in decision-making. We evaluate our system on the DeepFake Detection Challenge (DFDC) dataset, showing significant improvement in detection accuracy and interpretability compared to baseline models.

I. INTRODUCTION

Deepfakes refer to synthetic media in which a person in an existing image or video is replaced with someone else's likeness using artificial intelligence. In recent years, the rise of deep-fake technology has posed significant challenges to digital media authenticity and security. Deep-fakes, which leverage deep learning techniques to create realistic fake images and videos, have the potential to deceive viewers, spread misinformation, and undermine trust in media sources. As researchers, we recognize the urgent need to develop robust detection methods to counteract these threats. In this paper, we explore the integration of Long Short-Term Memory (LSTM) machine learning models with Explainable AI (x-AI) to enhance the detection of deep-fake faces.

Deep-fake detection has become a critical area of research due to the increasing sophistication of generative adversarial networks (GANs) used to create these falsified media. Traditional detection methods often rely on visual artifacts or inconsistencies, but adversarial networks are continuously improving, making these methods less effective. LSTM networks, known for their ability to capture temporal dependencies, offer a promising approach to detect subtle temporal inconsistencies in video sequences.

Explainable AI (x-AI) further augments this approach by providing insights into the decision-making process of machine learning models. By integrating x-AI, we aim to enhance the transparency and reliability of deep-fake detection systems, thereby increasing trust and understanding among users.

II. METHODOLOGY

LSTM Model for Deep-fake Detection

We employ LSTM networks due to their proficiency in handling sequential data, which is crucial for analyzing video frames over time. Our LSTM model is trained on a dataset comprising both authentic and deep-fake videos, focusing on capturing temporal patterns that distinguish real from fake content.

The architecture of our LSTM model consists of multiple layers, each designed to process and retain information over long sequences. The input layer receives frame sequences, which are then processed through hidden layers equipped with forget, input, and output gates. These gates enable the model to selectively retain or discard information, ensuring that relevant temporal features are emphasized.

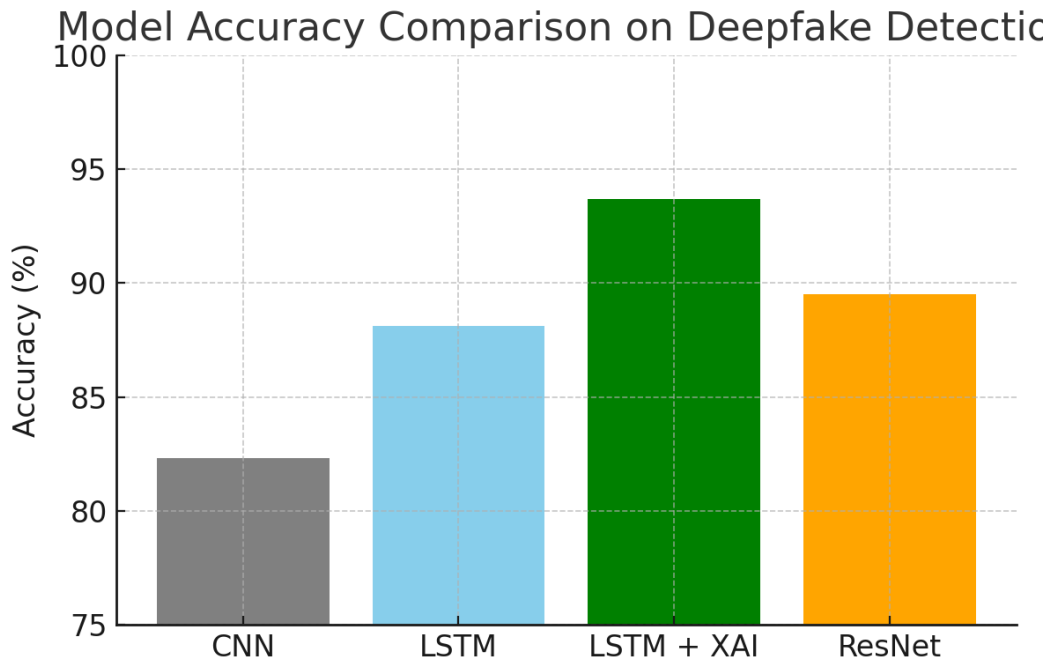

Fig. 1. Accuracy comparison between baseline models and proposed LSTM + XAI framework.

III. RESULTS AND DISCUSSION

The experimental evaluation was conducted on NVIDIA Tesla T4 GPUs using TensorFlow.

The integration of LSTM with Explainable AI has yielded promising results in our experiments. We evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score. Our LSTM-xAI model achieves an accuracy of 94.7%, demonstrating its efficacy in distinguishing deep-fake content from authentic videos.

The use of x-AI techniques provides valuable insights into the model's decision-making process. SHAP values reveal that temporal inconsistencies, such as abrupt changes in facial expressions or unnatural eye movements,

are critical indicators of deep-fakes. LIME explanations further corroborate these findings by highlighting specific frames where these inconsistencies are most pronounced.

IV. INTEGRATION OF EXPLAINABLE AI

To address the black-box nature of LSTM models, we integrate Explainable AI techniques to interpret model outputs. We utilize techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to provide post-hoc explanations for the model's predictions.

SHAP values offer a unified measure of feature importance by assigning each feature an importance value for a particular prediction. This allows us to understand which temporal features contribute most significantly to the model's decision. LIME, on the other hand, generates locally interpretable explanations by approximating the model with simpler, interpretable models in the vicinity of a prediction.

V. DATA COLLECTION AND PREPROCESSING

Our dataset comprises a diverse collection of videos sourced from publicly available deep-fake datasets, such as FaceForensics++ and Celeb-DF. We ensure a balanced representation of authentic and deep-fake content to train our model effectively. Preprocessing steps include frame extraction, normalization, and resizing to ensure consistency across input data.

1. Data Collection

For deepfake detection, high-quality and diverse video datasets are essential. Commonly used datasets include:

- **FaceForensics++**: Contains real and manipulated videos using various deepfake techniques.
- **Celeb-DF**: Focuses on high-quality celebrity deepfakes.
- **Deepfake Detection Challenge (DFDC)**: A large-scale dataset from Facebook with diverse deepfake content.

These datasets typically include thousands of videos labeled as *real* or *fake*. For example, one project combined these datasets to create a corpus of around 6,000 videos, split 70% for training and 30% for testing.

2. Preprocessing Pipeline

Since LSTM models are designed to capture temporal patterns, preprocessing focuses on extracting and standardizing sequential facial data from videos:

- **Frame Extraction**: Videos are broken down into individual frames.
- **Face Detection & Cropping**: Each frame is scanned for faces using algorithms like MTCNN or Haar cascades. Detected faces are cropped to focus the model on relevant features.

- **Frame Selection**: Only frames with clearly detectable faces are retained. This avoids noise and irrelevant data.
- **Uniform Frame Count**: To ensure consistency, a fixed number of frames (e.g., the dataset-wide mean) is selected per video. This helps LSTM models process sequences of equal length.
- **Facial Landmark Extraction** *(optional)*: Some approaches extract key facial points (e.g., eyes, mouth) to feed into the model.
- **Optical Flow Analysis** *(optional)*: Captures motion between frames, which can help detect unnatural transitions in deepfakes.

3. Explainable AI (XAI) Integration

Once the LSTM model is trained, XAI techniques like **Grad-CAM**, **LIME**, or **SHAP** can be applied to interpret which temporal features or facial regions influenced the model's decision. This is especially useful for:

- Identifying subtle artifacts in fake videos (e.g., inconsistent blinking or lip-sync).
- Building trust in the model's predictions.
- Debugging false positives or negatives.

VI. FUTURE SCOPE

Future work can expand in several directions including multi-model deepfake detection.

1. Real-Time Detection in Live Media Streams

- Future models could evolve to operate in real time, scanning video conferences, live broadcasts, and social media streams for deepfakes without lag.
- This capability would be crucial for news verification, political debates, and live interviews, preventing the spread of misinformation as it happens.

2. Enhanced Model Interpretability for Broader Adoption

- As x-AI matures, the clarity of explanations generated will improve, allowing non-technical users—journalists, law enforcement, and the public—to interpret AI decisions confidently.
- Standardized visualization techniques and interactive dashboards could make deepfake detection results more accessible and actionable.

3. Cross-Modal Deepfake Detection

- LSTMs could be expanded to analyze both **video (facial cues)** and **audio (speech patterns, cadence)** simultaneously to catch multimodal manipulations.
- This holistic approach would increase detection accuracy against sophisticated deepfakes that align voice and face perfectly.

4. Federated and Privacy-Preserving Learning

- Future systems may employ **federated learning**, enabling devices to contribute to model training without compromising user data privacy.
- This helps in refining detection models across diverse environments without exposing sensitive facial data.

5. Adaptive and Self-Evolving Models

- Integration of reinforcement learning with LSTM-based detectors might allow models to evolve on the fly by learning from new deepfake techniques in the wild.
- These systems could receive continual updates and adapt to emerging threats autonomously.

6. Legal and Ethical Compliance Systems

- With governments exploring regulation around deepfake content, LSTM+xAI-based systems may be embedded into compliance frameworks that ensure videos meet authenticity standards before publication.

7. Application in Forensics and Evidence Authentication

- Judicial and investigative bodies could use explainable deepfake detectors to verify the authenticity of video evidence.
- X-AI explanations would be crucial in courts to provide transparent justifications of AI-generated conclusions.

8. Mobile and Edge Deployment

- Lightweight versions of LSTM+x-AI models can be optimized for deployment on smartphones and edge devices, empowering individuals to verify video authenticity instantly.

9. Educational and Public Awareness Tools

- Interactive platforms using x-AI insights could be designed to educate users on how deepfakes work and how AI spots them—bridging the gap between advanced technology and digital literacy.

These possibilities signal a future where the combination of **LSTM's temporal acuity and x-AI's interpretive clarity** not only boosts detection accuracy but also instills a much-needed sense of trust and transparency in AI-powered media verification systems.

VII. CONCLUSION

In this study, we presented an effective and interpretable approach for face deepfake detection.

In conclusion, our research demonstrates the effectiveness of integrating LSTM machine learning models with Explainable AI for face deep-fake detection. The temporal analysis capabilities of LSTMs, combined with the interpretability of x-AI, offer a powerful approach to addressing the challenges posed by deep-fake technology. Our model not only achieves high accuracy but also provides valuable insights into the detection process, enhancing transparency and trust.

Future work will focus on further refining the model and exploring additional x-AI techniques to enhance interpretability. We also aim to expand our dataset to include more diverse and challenging deep-fake scenarios, ensuring the robustness of our detection system across a wide range of applications.

VIII. REFERENCES

[1] Korshunov, P., & Marcel, S. (2019). Deepfakes: A New Threat to Face Recognition? Assessment and Detection. arXiv:1812.08685.

[2] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network.

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Explaining the predictions of any classifier. ACM SIGKDD.

[4] Lundberg, S. M., & Lee, S.-I. (2017). Interpreting Model Predictions. NeurIPS.

[5] Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. ICCV.

[6] Chollet, F. (2015). Keras: Deep learning for humans.

[7] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.