Defakedev: Deep Fake Detector Integrated With Virtual Universe

Mr. Ashis Kumar Mishra[1] Saurabh Satapathy[2] Sumit Das Mohapatra[3] Omm Prakash Jena[4]
[1]Assistant Professor [2,3,4]Research Scholar
[1,2,3,4]Department of Computer Science
[1,2,3,4]Odisha University of Technology and Research, Bhubaneswar, Odisha, India

Abstract — Introducing a state-of-the-art deep learning method, our approach utilizes a Res-Next CNN for extracting frame-level features and an LSTM-based RNN for classification, effectively distinguishing between authentic and AI-generated fake videos. Trained on a diverse dataset comprising Face-Forensic++, Deepfake Detection Challenge, and Celeb-DF datasets, our method demonstrates robustness in identifying manipulated media. Furthermore, the integration of this deepfake detection model into a virtual universe allows for real-time detection of deepfakes, enriching user experiences within the simulated environment. This integration fosters greater awareness of the risks associated with manipulated media and offers an interactive platform for users to engage with and comprehend the implications of deepfake technology.

Keywords: Deep Learning, Deepfake Detection, Res-Next CNN, LSTM-based RNN, Face-Forensic++, Deepfake Detection Challenge, Celeb-DF, virtual universe, real-time detection, manipulated media, interactive platform

I. Introduction

In recent years, the proliferation of deepfake technology has raised significant concerns regarding its potential for misuse and manipulation in various domains, including politics, entertainment, and personal privacy. Deepfakes, hyper-realistic synthesized videos created using artificial intelligence techniques, have become increasingly accessible due to advancements in deep learning algorithms and the widespread availability of powerful computing resources. These deepfakes pose a threat to societal trust, as they can be used to fabricate events, spread misinformation, and exploit individuals through means such as revenge porn and blackmail. In response to these challenges, there is a pressing need for robust and effective methods to detect and combat the spread of deepfake content. In this study, we propose a novel deep learning-based approach aimed at accurately identifying deepfake videos. By harnessing the capabilities of Res-Next Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM)-based Recurrent Neural Networks (RNN) for classification, our method aims to differentiate between authentic and AI-generated fake videos. Additionally, to ensure the model's effectiveness in real-world scenarios, we integrate it into a virtual universe, enabling real-time detection of deepfakes as users engage with simulated environments. This integration not only enhances awareness of the dangers posed by manipulated media but also provides an interactive platform for users to actively engage with and understand the implications of deepfake technology. Through this interdisciplinary approach, we aim to contribute to the ongoing efforts to mitigate the risks associated with deepfake proliferation and safeguard the integrity of digital content.

II. Literature Review

Deepfake technology has garnered significant attention in recent years due to its potential to deceive and manipulate digital media content. Detecting and mitigating the spread of deepfakes is crucial to preserving the integrity of online information and combating the proliferation of misinformation. In this literature review, we examine existing research on deepfake detection techniques and their applications across various sectors. We summarize key findings from relevant studies, highlight challenges and advancements in the field, and discuss the significance of integrating deepfake detection into virtual environments like OVER Universe.

Authors and Date	Main Contribution	Key Finding	Limitations
Antipov, M., Baccouche, M., & Dugelay, J.-L. (2017)	Proposed a method for synthetic aging of human faces using acGAN	Identity-preserving approach preserves original identity	Focuses solely on face aging, may not address other manipulations
Korshunov, P., & Marcel, S. (2018)	Benchmarking systems for detecting audio-visual inconsistencies	LSTM-based systems detect tampering consistently	Focuses primarily on audio-visual inconsistencies
Korshunov, P., & Marcel, S. (2018)	Introduced a database of Deepfake videos and evaluated detection algorithms	VGG and Facenet vulnerable to Deepfakes	Dataset limitations may hinder detection effectiveness
Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019)	Created FaceForensics++ dataset and proposed a benchmark	Current manipulation methods detectable, transfer learning important	Dataset may not cover all manipulation techniques
Singh, B., & Sharma, D. K. (2021)	Proposed a multi-modal approach to detect fake images	Achieved high accuracy in detecting fake images	May not cover all types of fake images or consider textual content
Tariq, S., Abuadbba, A., & Moore, K. (2023)	Addressed security implications of Deepfakes in the metaverse	Deepfakes pose significant security threats	Focuses on security implications, may lack technical depth

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2020)	Developed real-time facial reenactment system, Face2Face	Enabled live animation with potential applications in VR/AR	System performance may be limited by input quality and complexity
Cunha Lacerda, G., & Vasconcelos, R. C. (2022)	Presented a method using EfficientNet and Celeb-DF dataset for deepfake detection	Achieved accuracy of more than 93%	Limited to images of 224x224 pixels
Agarwal, S., El-Gaaly, T., Farid, H., & Lim, S.-N. (2020)	Developed a technique leveraging facial and behavioral biometrics for detecting face-swap deepfakes	Less vulnerable to counterattack and generalizes well	May struggle to classify lip-sync deep fakes
Jiang, L., Li, R., Wu, W., & Loy, C. C. (2020)	Proposed a large-scale dataset for face forgery detection	Facilitates research on face forgery detection	Future works include expanding the dataset and improving evaluation metrics
Kingra, S., Aggarwal, N., & Kaur, N. (2022)	Analyzed visual media tampering detection techniques, with a focus on deepfake detection	Provided a critical summarization of state-of-the-art approaches	Detection methods provide a partial solution to the problem
Coccomini, D., Messina, N., Gennaro, C., & Falchi, F. (2023)	Demonstrated the effectiveness of mixed convolutional-transformer networks for deepfake detection	Used EfficientNet and Vision Transformers for state-of-the-art results	Future work may involve improving classification performance
Lai, Y., Luo, Z., & Yu, Z. (2023)	Introduced a framework for face forgery detection and localization using a Segment Anything Model	Proposed Multiscale Adapter and Reconstruction Guided Attention for robust forgery localization	Demonstrated effectiveness across different qualities of face images
Saha, S., Perera, R., Seneviratne, S., Malepathirana, T., Rasnayaka, S., Geethika, D., Sim, T., & Halgamuge, S. (2023)	Proposed a method for robust deepfake detection at frame, segment, and video levels	Achieved robust IoU metrics across single and multi-segment deepfakes	Limited to supervised pretraining of the image encoder
Muppalla, S., Jia, S., & Lyu, S. (2023)	Introduced a method for audio-visual deepfake detection focusing on multimodal inconsistency features	Demonstrated good adaptability across various feature extraction networks	Future work includes developing more robust multimodal networks

Table 1: Literature Survey

III. PROPOSED MODEL

Our proposed model for deepfake detection leverages the synergy between Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to tackle the complexities of identifying manipulated videos. CNNs, renowned for their proficiency in spatial feature extraction, are adept at analyzing individual frames of a video. We integrate a pre-trained ResNext CNN model into our architecture, harnessing its ability to capture intricate visual cues essential for discerning deepfakes effectively. By utilizing a pre-trained model, we optimize feature extraction, facilitating thorough analysis of each frame's content.

However, the challenge of deepfake detection extends beyond static frame analysis; it necessitates an understanding of temporal dependencies within video sequences. To address this, our model incorporates a Long Short-Term Memory (LSTM) network, a specialized form of RNN designed for sequence processing. By feeding the extracted features from the CNN into the LSTM, our model gains the ability to capture temporal dynamics and subtle changes across frames. This fusion enables our model to discern patterns indicative of deepfake manipulation, enhancing its overall accuracy and reliability.

Complementing the core CNN and LSTM components, our model architecture incorporates additional layers such as ReLU activation, dropout, and adaptive average pooling. These mechanisms play a pivotal role in enhancing model robustness and mitigating overfitting during training. By incorporating these techniques, we ensure that our model generalizes well to unseen data and remains effective in scenarios where distinguishing between real and fake videos poses a significant challenge. In essence, our proposed model amalgamates the strengths of CNNs and RNNs, bolstered by supplementary layers, to conduct comprehensive analysis and accurately identify deepfake videos with high precision.

A. DeFakeDV Architecutre

1) Preprocessing Details

The initial phase involves importing videos using the glob module and determining the mean number of frames per video using cv2.VideoCapture. Standardization is achieved by selecting 150 frames as the optimal value. Videos are then

split into frames and cropped to isolate facial regions. Utilizing VideoWriter, cropped frames are encoded into new videos with specifications of 112 x 112 pixels resolution and a frame rate of 30 fps. This preprocessing ensures uniformity and prepares the data for subsequent analysis.

2) Model Layers Details

The architecture encompasses a ResNext CNN pre-trained model, which offers 50 layers and a 32 x 4 dimensionality. A Sequential layer organizes feature vectors from the ResNext model, facilitating sequential passage to the LSTM layer. The LSTM layer, with 2048 latent dimensions and hidden layers, enables temporal sequence analysis. ReLU activation functions and a dropout layer with a 0.4 dropout rate mitigate overfitting. An Adaptive Average Pooling layer further refines feature extraction by reducing variance and computational complexity.

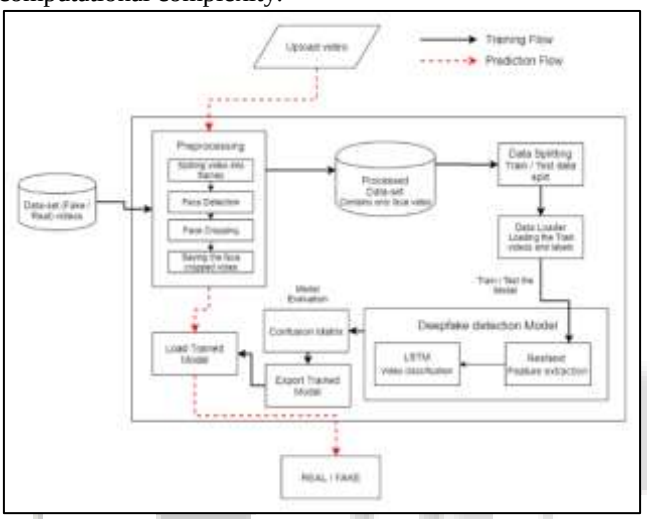

Fig. 1: DeFakeDV Architecture

3) Training and Evaluation Details

The dataset undergoes a balanced split into train and test sets, with a ratio of 70% for training and 30% for testing. Data loaders are utilized to load videos and their corresponding labels in batches of 4. Training is executed for 20 epochs using the Adam optimizer with a learning rate of 1e-5 and weight decay of 1e-3. Cross-entropy loss function is employed for classification, and a softmax layer with two output nodes (REAL or FAKE) provides prediction probabilities. Model performance is evaluated using confusion matrices to determine accuracy and validate the efficacy of the trained model.

IV. IMPLEMENTATION

The implementation of the proposed model encompasses a series of pivotal steps aimed at ensuring the robust detection of deepfake videos. Initially, exhaustive efforts are made to curate real and fake video datasets from a myriad of diverse sources, thereby ensuring an all-encompassing coverage spanning various video types. These meticulously curated datasets then undergo a rigorous preprocessing phase, wherein intricate facial regions are meticulously extracted and the video content is standardized to facilitate uniform analysis. This preprocessing stage is indispensable as it sets the foundation for subsequent phases, ensuring that the ensuing analysis is conducted with utmost precision and consistency.

Subsequent to the preprocessing stage, the videos are meticulously partitioned into distinct training and testing sets, a critical step that lays the groundwork for model development and evaluation. With the delineation of these sets, the model embarks on its training journey, harnessing an optimized amalgamation of hyperparameters, including but not limited to the learning rate and weight decay. This phase is marked by meticulous fine-tuning, where the model iteratively refines its parameters to attain a heightened level of efficacy in discerning between authentic and forged videos. Upon successful completion of the training regimen, the model is poised for deployment within an application environment, thereby facilitating real-time predictions concerning the authenticity of videos.

An innovative application scenario emerges wherein the model finds integration into a virtual universe actualized through cutting-edge A-Frame technology. This integration heralds a paradigm shift in user engagement, wherein individuals are afforded the unique opportunity to dynamically detect deepfake videos and images within a simulated environment. This immersive experience not only augments awareness regarding the perils posed by deepfake technology but also furnishes users with an interactive platform to comprehend the far-reaching ramifications of manipulated media. Thus, the seamless fusion of advanced detection mechanisms with immersive virtual environments not only fortifies our defenses against digital deception but also serves as a conduit for fostering greater societal resilience in the face of evolving technological threats.

Fig. 2: DeFakeDv

V. USE CASES

A. Social Media Content Verification:

Social media platforms can utilize the deepfake detection model to verify the authenticity of user-generated content, including images and videos shared by users. By integrating the model into their content moderation systems, social media platforms can automatically analyze multimedia content before it is published, identifying and flagging deepfake content to prevent its spread and maintain the platform's credibility. This proactive approach enhances user safety and protects individuals from misinformation, cyberbullying, and identity theft, fostering a more trustworthy online environment.

Fig. 3: Social Media Content verification by DeFakeD

B. News Verification and Fact-Checking:

News agencies can integrate the deepfake detection model into their content management systems to automatically analyze multimedia content before publishing. By doing so, they ensure that news articles and reports are based on authentic information, enhancing the reliability and trustworthiness of their content. Identifying and filtering out deepfakes from news content allows agencies to mitigate the spread of false information and uphold journalistic integrity, ultimately contributing to a more informed society.

Fig. 4: News verification & Fact Checking by DeFakeD

C. Intellectual Property Protection:

Production studios and streaming platforms can utilize the deepfake detection model to detect unauthorized deepfakes of copyrighted material. By identifying deepfakes impersonating celebrities or public figures, the entertainment industry can protect intellectual property rights and prevent the spread of pirated or manipulated content. This proactive approach safeguards the reputation and image of talent while maintaining the integrity of creative works, ensuring compensation for content creators.

Fig. 5: Intelectal Property Protection by DeFakeD

D. Legal Evidence Authentication:

Legal professionals can integrate the deepfake detection model into their systems to authenticate multimedia evidence presented in court proceedings. By ensuring that evidence submitted is genuine and admissible, the model enhances the reliability of legal proceedings and contributes to fair and just outcomes. Law enforcement agencies can also utilize the model to identify deepfake content used for fraudulent purposes, such as creating false alibis or tampering with surveillance footage, strengthening the integrity of criminal investigations and maintaining public trust in the justice system.

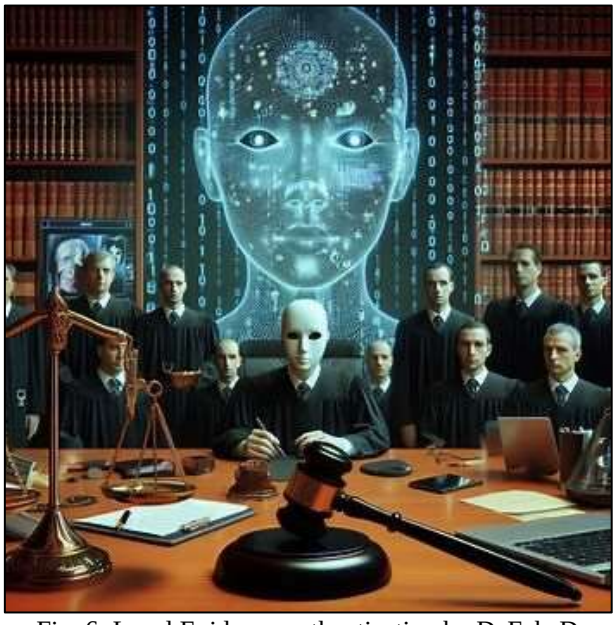

Fig. 6: Legal Evidence authentication by DeFakeD

E. Others:

Organizations across various sectors can employ the deepfake detection model for cybersecurity measures, identity verification, educational initiatives, and healthcare applications. In cybersecurity and identity verification, the model aids in identifying and mitigating risks associated with manipulated media, protecting sensitive data from phishing attacks, identity theft, and social engineering scams. Educational institutions and advocacy groups utilize the model to promote media literacy and critical thinking skills, raising awareness about the dangers of manipulated media and empowering individuals to discern real from fake content. Furthermore, in healthcare applications, the model ensures the security and integrity of patient data in telemedicine and remote healthcare settings, verifying the authenticity of medical images and videos to prevent the dissemination of false information and safeguard patient confidentiality.

Fig. 6: Other Implementations of DeFakeD

VI. CONCLUSION

In conclusion, our proposed model offers a robust solution for detecting deepfake videos by combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Through the integration of a pre-trained ResNext CNN model and an LSTM network, we achieve accurate identification of spatial and temporal features crucial for distinguishing authentic from manipulated content. Supplementary layers ensure model robustness and prevent overfitting, ensuring reliable performance across diverse datasets. By seamlessly integrating our model into a virtual universe, we provide real-time detection capabilities, enhancing awareness of the risks associated with deepfakes. Overall, our approach represents a significant advancement in deepfake detection, promising to mitigate the spread of misinformation and safeguard digital integrity.

REFERENCES

[1] Antipov, M., Baccouche, M., & Dugelay, J.-L. (2017). Face aging with conditional generative adversarial networks. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 2089-2093).
[2] Korshunov, P., & Marcel, S. (2018). Speaker inconsistency detection in tampered video. In European Signal Processing Conference (EUSIPCO).
[3] Korshunov, P., & Marcel, S. (2018). DeepFakes: a New Threat to Face Recognition? Assessment and Detection.
[4] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. In 2019 IEEE International Conference on Image Processing (ICIP).
[5] Singh, B., & Sharma, D. K. (2021). Predicting image credibility in fake news over social media using multi-modal approach. Neural Computing and Applications, 34(8), 21503-21517.
[6] Tariq, S., Abuadbba, A., & Moore, K. (2023). Deepfake in the Metaverse: Security Implications for Virtual Gaming, Meetings, and Offices.
[7] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2020). Face2Face: Real-time Face Capture and Reenactment of RGB Videos.
[8] Cunha Lacerda, G., & Vasconcelos, R. C. (2022). A Machine Learning Approach for DeepFake Detection. arXiv:2209.13792v1.
[9] Agarwal, S., El-Gaaly, T., Farid, H., & Lim, S.-N. (2020). Detecting Deep-Fake Videos from Appearance and Behavior. arXiv:2004.14491v1.
[10] Jiang, L., Li, R., Wu, W., & Loy, C. C. (2020). DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. arXiv:2001.03024v2.
[11] Kingra, S., Aggarwal, N., & Kaur, N. (2022). Emergence of deepfakes and video tampering detection approaches: A survey.
[12] Coccomini, D., Messina, N., Gennaro, C., & Falchi, F. (2023). Combining EfficientNet and Vision Transformers for Video Deepfake Detection. arXiv:2305.06564v4.
[13] Lai, Y., Luo, Z., & Yu, Z. (2023). Detect Any Deepfakes: Segment Anything Meets Face Forgery Detection and Localization. arXiv:2306.17075v1.
[14] Saha, S., Perera, R., Seneviratne, S., Malepathirana, T., Rasnayaka, S., Geethika, D., Sim, T., & Halgamuge, S. (2023). Undercover Deepfakes: Detecting Fake Segments in Videos. arXiv:2305.06564v4.
[15] Muppalla, S., Jia, S., & Lyu, S. (2023). Integrat?ng Audio-Visual Features For Mult?modal Deepfake Detect?on. arXiv:2310.03827v1.

