Unmasking the Unreal: Multi-Modal Deepfake Video Detection

Mrs. Nita J. Mahale[1], Tejas A. Kulkarni[2], Paras V. Yadav[3] , Ruturaj T. Mahekar[4], Harshal K. Bandgar[5]
[1]*Assistant Professor, Department of Artificial Intelligenece and Data Science, D. Y. Patil College of Engineering, Pune*
[2,3,4,5]*Student, Department of Artificial Intelligenece and Data Science, D. Y. Patil College of Engineering, Pune*

***Abstract*—Deepfake technology poses a significant threat to digital media authenticity and trust. This paper presents a multi-modal deep learning framework for detecting deepfake videos by integrating three complementary detection mechanisms: Convolutional Neural Networks (CNN) for spatial artifact detection, Bidirectional Long Short-Term Memory (BiLSTM) networks for temporal pattern analysis, and eye-blink frequency analysis for physiological authenticity verification. The proposed system processes video frames through a CNN architecture to identify facial inconsistencies, employs BiLSTM to capture temporal anomalies across frame sequences, and leverages OpenCV-based eye-blink detection to assess natural human behavior patterns. The final classification is performed through weighted majority voting, combining predictions from all three modalities. Experiments conducted on the CelebDF dataset demonstrate the effectiveness of the multi-modal approach, achieving improved detection accuracy compared to single-modality methods. The system utilizes 500 real and 800 fake video samples for training, with MTCNN for face detection and Haar cascades for eye detection. Results indicate that the ensemble approach provides robust deepfake detection capabilities, addressing limitations inherent in individual detection techniques**

***Index Terms*—Deepfake Detection, Convolutional Neural Networks, BiLSTM, Eye-Blink Analysis, Multi-Modal Learning, Video Forensics.**

I. INTRODUCTION

Deepfake technology, powered by advanced generative adversarial networks (GANs) and deep learning algorithms, has enabled the creation of highly realistic synthetic media that is increasingly difficult to distinguish from authentic content. While these technologies offer creative possibilities, they also present serious challenges to information integrity, personal privacy, and security. The proliferation of deepfake videos across social media platforms has raised concerns about their potential use in misinformation campaigns, identity theft, and fraud.Traditional video authentication methods rely primarily on metadata analysis or compression artifacts, which can be easily manipulated or removed. As deepfake generation techniques become more sophisticated, there is an urgent need for robust detection systems capable of identifying synthetic content through multiple analysis dimensions.This research addresses the deepfake detection challenge by proposing a multi-modal detection framework that simultaneously analyzes spatial, temporal, and physiological characteristics of video content. The proposed system integrates three complementary approaches: (1) CNN-based spatial feature extraction from cropped facial regions to identify visual artifacts, (2) BiLSTM-based temporal sequence analysis to detect inconsistencies in frame transitions, and (3) eye-blink frequency analysis to verify natural human behavioral patterns. By combining these modalities through weighted majority voting, the system achieves enhanced detection performance while maintaining computational efficiency.The primary contributions of this work include: (i) a novel multi-modal architecture combining spatial, temporal, and physiological analysis, (ii) implementation of an eye-blink detection mechanism using OpenCV Haar cascades for compatibility and efficiency, (iii) empirical validation on the CelebDF dataset demonstrating improved accuracy over single-modality approaches, and (iv) a practical framework suitable for real-world

deployment with moderate computational requirements.

II. LITERATURE REVIEW

Deepfake detection has emerged as an active research area, with numerous approaches proposed to address the evolving threat. Early detection methods primarily focused on spatial artifacts introduced during the face-swapping process. Li et al. (2020) demonstrated that CNNs could identify subtle inconsistencies in facial regions, particularly around boundaries and texture patterns. Their work established that deep learning models trained on face crops could effectively distinguish between real and synthetic faces.Temporal analysis approaches gained prominence as researchers recognized that deepfake generation often introduces temporal inconsistencies not present in authentic videos. Sabir et al. (2019) proposed using recurrent neural networks to analyze frame sequences, showing that temporal artifacts could be detected through sequence modeling. The use of LSTM and its bidirectional variants (BiLSTM) has proven effective in capturing temporal dependencies that reveal manipulation artifacts.Physiological signal analysis represents another promising direction. Li et al. (2018) explored the use of eye-blink patterns as a biometric signature, noting that deepfake videos often exhibit unnatural or reduced blink frequencies due to limitations in face reenactment algorithms. This approach leverages the fact that natural human blinking follows specific patterns that are difficult to replicate accurately in synthetic content.Ensemble methods combining multiple detection modalities have shown superior performance. Zhou et al. (2020) proposed a multi-stream framework that integrates spatial and temporal features, achieving state-of-the-art results on benchmark datasets. However, their approach required complex architectures with high computational overhead.Recent work has emphasized the importance of practical deployment considerations. The CelebDF dataset, introduced by Li et al. (2020), provides a challenging benchmark with higher quality deepfakes compared to earlier datasets. This dataset has become a standard evaluation platform for comparing detection methods. Our approach builds upon these foundations by integrating spatial CNN analysis, temporal BiLSTM modeling, and eye-blink frequency assessment into a unified framework with efficient implementation suitable for practical applications.

III. METHODOLOGY

3.1 Dataset

The proposed system is evaluated on the CelebDF dataset, which consists of real celebrity interview videos and corresponding deepfake versions generated using advanced face-swapping techniques. For this study, we utilize a subset of 500 real videos and 800 fake videos to balance computational efficiency with sufficient training data. Videos are processed at their original resolutions, with frames extracted using uniform sampling across the video duration.

3.2 Preprocessing

Video preprocessing involves frame extraction, face detection, and region cropping. For each video, we extract 10 evenly-spaced frames to ensure representative temporal coverage while maintaining computational tractability. Face detection is performed using MTCNN (Multi-task Cascaded Convolutional Networks), which provides robust face localization even under varying lighting and pose conditions. Detected faces are cropped and resized to 224×224 pixels for CNN input, and frames are resized to 128×128 pixels for BiLSTM sequence processing.

3.3 CNN-Based Spatial Detection

The CNN component is designed to identify spatial artifacts in facial regions. The architecture consists of three convolutional blocks, each followed by max-pooling layers. The first block contains 32 filters (3×3 kernel), the second contains 64 filters, and the third contains 128 filters. After flattening, two fully connected layers (128 and 1 neurons) with dropout (0.5) are applied. The output layer uses a sigmoid activation function to produce binary classification probabilities. Data augmentation including horizontal flipping, zoom variations, and slight rotations is applied during training to improve generalization. The model is trained for 5 epochs using the Adam optimizer with binary cross-entropy loss.

3.4 BiLSTM-Based Temporal Detection

The temporal analysis module employs a Bidirectional LSTM architecture to capture sequential patterns across video frames. The input consists of sequences

of 10 frames (128×128×3) extracted from each video. A TimeDistributed layer applies convolutional operations to each frame, extracting spatial features before temporal modeling. Two TimeDistributed convolutional blocks (32 and 64 filters) with max-pooling are followed by flattening and a Bidirectional LSTM layer with 128 units. Dropout layers (0.5 and 0.3) prevent overfitting, and a dense layer with sigmoid activation provides the final classification. The BiLSTM model is trained for 5 epochs with a batch size of 8.

3.5 Eye-Blink Analysis

Eye-blink detection serves as a physiological authenticity indicator. Using OpenCV Haar cascades for face and eye detection, the system analyzes eye aspect ratio (EAR) across sampled frames. The EAR is computed as the ratio of eye region height to width, with values below a threshold (0.25) indicating closed eyes. Blink events are identified when consecutive frames show eyes closed, followed by reopening. The blink frequency is compared against expected natural patterns (typically 15-20 blinks per minute), with deviations suggesting potential deepfake content. This heuristic approach converts blink count into a fake probability estimate, providing an additional detection signal without requiring training.

3.6 Ensemble Classification

The final classification combines predictions from all three modalities through weighted majority voting. Each model provides a fake probability score, and the ensemble computes a weighted average based on individual model confidence. Additionally, a simple majority vote is performed on binary labels. The final decision incorporates both approaches, with the weighted method providing probability estimates and the majority vote offering a discrete classification. This dual approach ensures robustness against individual model failures while maintaining interpretability.

IV. RESULTS AND DISCUSSION

4.1 Experimental Setup

The dataset is split into training (80%) and testing (20%) sets with stratified sampling to maintain class balance. All models are implemented using TensorFlow/Keras, with training performed on standard hardware configurations. The evaluation metrics include training accuracy, test accuracy, and prediction confidence scores for individual models and the ensemble.

4.2 Individual Model Performance

The CNN-based face detection model achieves competitive performance in identifying spatial artifacts. Training on cropped facial regions allows the model to focus on facial inconsistencies introduced during deepfake generation. The BiLSTM temporal model demonstrates effectiveness in capturing frame-to-frame inconsistencies, particularly useful for detecting subtle temporal artifacts that may not be apparent in single-frame analysis.The eye-blink analysis component provides complementary information, leveraging physiological patterns that are inherently difficult to synthesize naturally. While this method alone may not achieve high accuracy, it contributes valuable additional evidence to the ensemble decision.

4.3 Ensemble Performance

The multi-modal ensemble approach demonstrates improved performance compared to individual modalities. By combining spatial, temporal, and physiological analysis, the system achieves higher accuracy and greater robustness against adversarial examples. The weighted voting mechanism effectively balances contributions from each modality, with models exhibiting higher confidence receiving greater weight in the final decision.The system successfully identifies deepfake content with improved confidence levels, addressing the challenge of false positives and false negatives that plague single-modality approaches. The integration of eye-blink analysis adds a novel dimension to deepfake detection, providing an additional verification mechanism that is difficult for generative models to replicate accurately.

4.4 Limitations and Future Work

Several limitations should be acknowledged. The current implementation uses a subset of the CelebDF dataset; expanding to the full dataset would improve model generalization. The eye-blink detection relies on heuristic thresholds that may need adjustment for different video conditions. Computational efficiency could be enhanced through model optimization and hardware acceleration

.Future work will explore: (i) integration of audio analysis for multi-modal enhancement, (ii) adaptation to newer deepfake generation techniques, (iii) real-time detection capabilities for streaming applications, (iv) explainability mechanisms to provide interpretable detection results, and (v) evaluation on additional benchmark datasets to assess cross-dataset generalization.

V. CONCLUSION

This paper presents a multi-modal deep learning framework for deepfake video detection that integrates CNN-based spatial analysis, BiLSTM temporal modeling, and eye-blink frequency assessment. The ensemble approach demonstrates improved detection accuracy by leveraging complementary information from multiple analysis dimensions. Experimental results on the CelebDF dataset validate the effectiveness of the proposed method, showing that combining spatial, temporal, and physiological signals provides robust detection capabilities.The practical implementation using MTCNN for face detection and OpenCV for eye-blink analysis ensures compatibility and efficiency, making the system suitable for real-world deployment. The weighted majority voting mechanism effectively combines individual model predictions, providing both probability estimates and discrete classifications.As deepfake generation techniques continue to evolve, multi-modal detection approaches that analyze content from multiple perspectives will remain essential for maintaining media authenticity. The framework presented here contributes to this effort by demonstrating how spatial, temporal, and physiological analysis can be integrated into a unified detection system. Future enhancements focusing on real-time processing, cross-dataset generalization, and explainability will further strengthen the practical utility of deepfake detection systems.

REFERENCES

[1] " Deep Learning for Deepfakes Creation and Detection: A Survey": Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec, Saeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyeng

[2] A Contemporary Survey on Deepfake Detection: Datasets, Algorithms" Liang Yu Gong *,† andXueJunLi*,† Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland 1010, New Zealan

[3] "Deepfake video detection: challenges and opportunities": Achhardeep Kaur1 · Azadeh Noori Hoshyar2 · Vidya Saikrishna3 · Selena Firmin1 · Feng Xia4 Artificial Intelligence Review (2024) 57:159

[4] "Deepfake detection using deep learning methods: A systematic and comprehensive review": ArashHeidari1 | Nima Jafari Navimipour2,3 | HasanDag4 | MehmetUnal5 DOI:10.1002/widm.1520

[5] "Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks": 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) | 978-1-6654-6835-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/HORA55278.2022.9799858

[6] "The DeepFake Detection Challenge (DFDC) Dataset": Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, Cristian Canton Ferrer Facebook AI arXiv:2006.07397v4 [cs.CV] 28 Oct 2020

[7] "Implementation of a Deepfake Detection System using Convolutional Neural Networks and Adversarial Training" : 2023 3rd International Conference on Intelligent Technologies (CONIT) | 979-8-3503-3860-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CONIT59222.2023.10205614

[8] "Deepfake Detection: Leveraging InceptionResNetV2 and LSTM for Enhanced Accuracy ": 025 International Conference on Pervasive Computational Technologies (ICPCT) | 979-8-3315-0868-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICPCT64145.2025.10941494

[9] "Generation And Detection of Deepfakes using Generative Adversarial Networks (GANs) and Affine Transformation": 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-

5/23/$31.00 ©2023 IEEE | DOI: 1109/ICCCNT56998.2023.10307811

[10] "Deep Learning Based Deepfake Video Detection System": 2025 3rd International Conference on Disruptive Technologies (ICDT) | 979-8-3315-1958-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDT63985.2025.10986738

[11] "Exposing Lip-syncing Deepfakes from Mouth Inconsistencies": 2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687902

[12] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[13] Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), 3(1), 80-87.

[14] Li, Y., Chang, M. C., & Lyu, S. (2018). In ictu oculi: Exposing AI generated fake videos by detecting eye blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS).

[15] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2020). Two-stream neural networks for tampered face detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops.

[16] Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. 2019 IEEE Winter Applications of Computer Vision Workshops.