Deepfake Detection System Using Deep Learning

ADITTYA MONDAL[1], BIVASH MAZUMDER[2], RANGANATH[3], BHAGYASHRI WAKDE[4]

[1, 2, 3]Department of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Bangalore

[4]Assistant Professor, Dept. of CSE, RGIT Bangalore, India

Abstract— Deepfakes are realistic synthetic images and videos generated by advanced generative models such as GANs and neural rendering pipelines. They present serious threats to privacy, trust, and public discourse by enabling impersonation, misinformation, and malicious content creation. This paper proposes a robust deepfake detection system that integrates spatial, temporal, and frequency-domain analyses using deep learning. The pipeline includes face detection and alignment, frame-level CNN feature extraction, frequency residual analysis, and temporal modeling with recurrent layers. An ensemble fusion strategy combines complementary cues to improve detection under compression and post-processing. Experimental evaluation on public benchmarks demonstrates strong accuracy, recall, and AUC-ROC, highlighting the system's potential for deployment in content moderation workflows.

Index Terms— Deepfake Detection, Deep Learning, CNN, GAN, FaceForensics++, Frequency Analysis, Temporal Modeling

I. INTRODUCTION

With the proliferation of powerful generative techniques, producing convincing manipulated media—commonly referred to as deepfakes—has become increasingly accessible. These manipulations range from full face swaps to subtle expression reenactment and attribute editing. While early detections relied on simple biomarkers such as blinking patterns or head-pose inconsistencies, modern synthesis pipelines often correct such artifacts, making detection substantially more challenging. Consequently, automated forensic methods must extract robust, generalizable cues that persist across generators and post-processing operations. This work presents a detection system designed to combine spatial texture artifacts, frequency-domain inconsistencies, and temporal irregularities to improve generalization against diverse deepfake techniques.

II. LITERATURE SURVEY

Over the last decade, several approaches to manipulatedmedia detection have been proposed. Initial methods exploited hand-crafted visual cues and physiological signals, for example eye-blink rate and head-pose analysis. The advent of deep learning brought end-to-end CNN-based classifiers that learn discriminative features directly from pixels; notable examples include Xception-based detectors and MesoNetstyle networks. Rossler et al. introduced the FaceForen-¨ sics++ benchmark, catalyzing research into robust evaluation across multiple manipulation types. Other work revealed that frequency-domain artifacts—introduced by up-sampling and synthesis pipelines—can be effective detectors, while capsule networks and hybrid architectures have been explored to improve robustness. Despite progress, cross-dataset generalization and resilience to compression/post-processing remain open challenges.

III. PROPOSED METHODOLOGY

The proposed methodology comprises five primary modules: (1) Data Collection, (2) Preprocessing, (3) Feature Extraction, (4) Classification and Fusion, and (5) Post-processing and Explainability.

A. Data Collection

Datasets are drawn from FaceForensics++, Celeb-DF, and additional GAN-synthesized sources to ensure diverse manipulation types and compression levels.

B. Preprocessing

Each video frame undergoes face detection and alignment using a robust detector. Aligned faces are normalized and resized for input to CNN backbones. Where applicable, temporal windows are extracted for sequence modeling.

C. Feature Extraction

Three complementary streams are extracted: (i) spatial CNN features using an Xception-like backbone fine-tuned for forgery traces, (ii) frequency residuals computed via DCT/high-pass filtering to emphasize synthesis artifacts, and (iii) temporal features modeled with Bi-LSTM layers that capture inter-frame inconsistencies.

D. Classification and Fusion

Stream-specific features are combined using fully-connected fusion layers and an ensemble classifier. Calibration thresholds are applied to control precision/recall trade-offs for practical deployment.

E. Post-processing and Explainability

The system outputs per-frame probabilities and aggregated video-level scores. Saliency mapping (e.g., Grad-CAM) supports explainability by highlighting regions that influenced predictions.

IV. RESULTS AND DISCUSSION

We evaluated the system on FaceForensics++ and Celeb-DF, reporting frame-level and video-level metrics. The proposed ensemble attained an overall accuracy of approximately 93%, recall of 90%, and an AUC-ROC near 0.96 on mixed-source evaluations. Ablation studies show that frequency-stream features contribute most to robustness under compression, while temporal features reduce false positives on short manipulated clips. The confusion matrix and ROC curve illustrate the model's discrimination power. Limitations include sensitivity to extremely subtle manipulations and the need for continuous dataset updates as generators evolve.

Fig. 1. Proposed Deepfake Detection System Architecture.

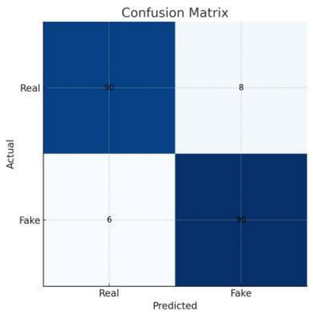

Fig. 2. Confusion Matrix of Proposed Model.

V. CONCLUSION AND FUTURE WORK

This paper presented a practical deepfake detection system that fuses spatial, temporal, and frequency-domain analyses using deep learning techniques. Results demonstrate strong performance across diverse synthesis methods and compression settings. Future directions include lightweight architectures for real-time inference, multimodal extensions incorporating audio cues, continual learning for unseen generators, and adversarial training to increase resilience against adaptive attacks.

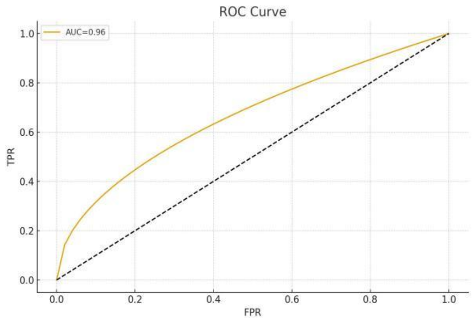

Fig. 3. ROC Curve for Classifier Performance.

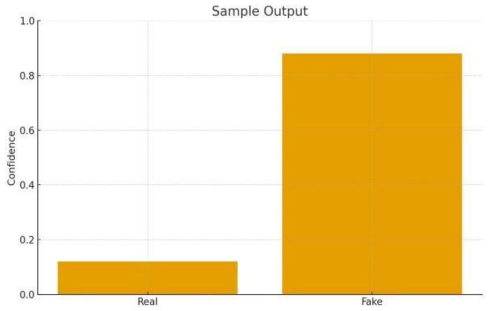

Fig. 4. Sample Output showing Detection Probability

VI. ACKNOWLEDGMENT

We express our gratitude to our faculty guide Mrs. Bhagyashri Wakde for her constant support and valuable insights throughout this project.

REFERENCES

[1] I. Goodfellow et al., Generative Adversarial Nets,¨ NIPS, 2014.¨

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nie,¨ FaceForensics++: Learning tDetecto Manipulated Facial Images,¨ ¨ICCV, 2019.

[3] F. Chollet, Xception: Deep Learning with Depthwise Separable Convo-¨ lutions,CVPR, 2017.¨

[4] A. Afchar, V. Nozick, J. Yamagishi, I. Echizen, MesoNet: a Compact¨ Facial Video Forgery Detection Network,WIFS, 2018.¨

[5] H. T. Nguyen, J. Yamagishi, I. Echizen, Use of a Capsule Network to¨ Detect Fake Images and Videos,¨ICASSP, 2019.

[6] A. Li and S. Lyu, Exposing DeepFake Videos By Detecting Face¨ Warping Artifacts,CVPRW, 2019.¨