A Multi-Modal Deep Learning Framework for Identifying Deepfakes and Synthetic Media Using Visual Forensics

Dr. Shiva Kumar K B Professor & Head, Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bangalore, India

Mr. Ranjeet Kumar Associate Professor, Dept. of Computer Science and Engineering, Don Bosco Insti- tute of Technology, Bangalore, India

Shivani R, Dept of CSE,Don Bosco Institute of Technology Vaishnavi Krishna, Dept of CSE,Don Bosco Institute of Technology Vaishnavi Rai, Dept of CSE,Don Bosco Institute of Technology

Abstract

With deepfakes becoming more convincing and widespread, it's getting harder to tell what's real and what's fake in digital media. This paper introduces a practical framework designed to spot deepfakes and synthetic visuals—whether they're images or videos. The approach blends different techniques: we look at subtle biometric signals like heartbeat rhythms through photoplethysmog- raphy (PPG), and we use deep learning models such as CNNs and LSTMs to track patterns over time. By combining these elements, the system becomes better at catching even well-made fakes. Our results show improved accuracy and better resistance to manipulation compared to tradition- al methods. This makes the framework useful for verifying media, supporting online safety, and helping digital forensics teams deal with the growing threat of fake content.

Keywords

Deepfake and Synthetic Media Detection, Visual Forensics Techniques, CNN and LSTM Mod- els, Photoplethysmography (PPG) Signals, Biometric Data Security, FaceForensics++ Dataset, Temporal and Spatial Feature Analysis, Robustness Against Deepfake Attacks, Advanced Deep Learning Architectures, CS-CNN and V4D Networks, Multimedia Content Verification

Introduction

Piracy In our fast-paced digital age, where media is evolving at lightning speed and artificial intelli- gence is woven into nearly every tech facet, it's becoming trickier to determine if the content we encounter is authentic. One of the most alarming developments in this arena is deepfake technology. Deepfakes leverage sophisticated AI algorithms to produce strikingly realistic yet entirely fabricat- ed videos, images, and audio. These can range from minor tweaks in facial expressions to entirely invented individuals who never existed. The widespread availability of deepfake tools and their in- creasing realism present serious risks.

They can fuel misinformation, invade privacy, and even jeopardize political systems and cybersecu- rity. The misuse of synthetic media is a pressing issue, making the creation of effective detection systems more crucial than ever. To tackle this challenge, researchers are blending digital forensics with AI techniques to identify signs of manipulation. However, as deepfake creation methods ad- vance, traditional detection tools often lag behind. In light of this, our study introduces a fresh ap- proach that merges physiological data with computational methods to enhance the accuracy and re- liability of deepfake detection. A pivotal aspect of our method involves Photoplethysmography (PPG)—a medical technique that tracks changes in blood volume

through the skin. While it may seem unrelated to video analysis, deepfake videos frequently struggle to replicate these subtle phys- iological signals, such as pulse-related color shifts in a person's face. By combining PPG signals with cutting-edge deep learning models—specifically Convolutional Neural Networks (CNNs) for detecting spatial patterns and Long Short-Term Memory (LSTM) networks for monitoring changes over time—our framework is crafted to catch even the faintest signs of manipulation. It pinpoints both visual and physiological discrepancies, aiding in the differentiation between genuine content and fakes. This research aspires to bolster the ongoing battle against deepfake technology and its implications.

Literature Review

[1] M.S. Khan and colleagues proposed a hybrid deepfake detection system that cleverly integrates sev- eral techniques: Convolutional Neural Networks (CNNs) for visual feature extraction, deep learning for complex pattern recognition, Content-Based Image Retrieval (CBIR) to compare image content effec- tively, K-Nearest Neighbors (KNN) for classification based on proximity in feature space, and Support Vector Machines (SVM) to draw decision boundaries. This multi-layered approach leverages the strengths of each method, resulting in a robust system that achieved an accuracy of 95.38%. Their work particularly stands out in handling biometric images, showing potential for secure identity verification systems.

[2] Chitra and Rajalakshmi (2023) designed a hybrid model combining CNNs with Long Short-Term Memory (LSTM) networks to detect deepfakes in videos by focusing on spatial details and temporal changes. The CNN extracts frame-level features, while the LSTM captures temporal dependencies across frames, such as unnatural motions or inconsistencies over time. Using frame differencing, the model highlights subtle changes between frames that often indicate tampering. This method achieved 93.12% accuracy, demonstrating effectiveness and feasibility for real-time biometric video authentica- tion in security systems.

[3] Al-Sabaawi et al. developed a sophisticated deepfake detection framework using XceptionNet, a deep CNN architecture known for efficient feature extraction, combined with Capsule Networks that preserve spatial hierarchies in the data. They enhanced this system further through transfer learning, uti- lizing pre-trained models to speed up learning on limited datasets. Tested extensively on the FaceForen- sics++ benchmark, their approach captured subtle facial cues and spatial inconsistencies that are com- mon in fake videos, achieving a solid accuracy of 94.08%. This method is promising for practical de- ployment in biometric identity verification.

[4] Vinayakumar R. and his team proposed a CNN-based deepfake detector enhanced with adversarial training and a GAN discriminator. The adversarial training involves exposing the model to challenging fake samples generated by a Generative Adversarial Network (GAN), which improves the model's ro- bustness against sophisticated forgeries. While their accuracy was slightly lower at 91.25%, the key con- tribution lies in improved resilience to adversarial attacks, making this system valuable in environments where attackers continuously improve their fake generation techniques.

[5] Andrii Bezverkhyi and co-authors focused on physiological biometrics such as unnatural eye blink- ing and facial asymmetry to detect deepfakes. Using classical neural networks, their model identifies subtle behavioral and biometric cues that are often overlooked in pixel-level analysis. Achieving an ac- curacy of 89.60%, their method emphasizes interpretability, allowing forensic analysts to understand why a video or image is flagged, which is crucial in legal and security contexts.

[6] N.V. Issac et al. (2023) introduced a deepfake detection framework combining CNN and LSTM net- works to capture both spatial features from individual frames and temporal patterns across video se- quences. This hybrid model excels at detecting manipulated videos by analyzing subtle frame-by-frame inconsistencies that alone would be hard to detect. Their system improves adaptability across different deepfake types, providing a versatile tool for biometric authentication where video content integrity is critical.

[7] Bommareddy, Samyal, and Dahiya (2023) proposed a system that pairs CNN architectures with ad- versarial training techniques to improve generalization and robustness of deepfake detectors. Their ap- proach trains the model on a wide variety of fake samples, including unseen ones, aiming to prevent overfitting and improve detection of novel deepfake methods. This approach is particularly useful for staying ahead of rapidly evolving deepfake technologies in biometric security.

[8] Dudykevych et al. (2022) designed a neural network-based detection system focusing on subtle bio- metric image manipulations, such as minor alterations to facial features or textures that are hard to spot with traditional methods. Their research underscores the increasing vulnerability of biometric data to so- phisticated attacks and stresses the need for forensic-grade detection tools with high precision and low false positive rates.

[9] Jain and Varshney (2024) performed a comprehensive analysis of machine learning models applied to content-based image

retrieval (CBIR), demonstrating how this technique can assist in deepfake detec- tion. By comparing suspect images against large databases of authentic visuals, CBIR helps identify ma- nipulated content based on feature similarity metrics. Their study highlights the potential of integrating CBIR with other machine learning classifiers to build robust biometric verification systems.

[10] Chauhan et al. (2022) conducted an extensive evaluation of multiple deep learning architectures and datasets to benchmark performance in deepfake detection. Their work systematically compared models on metrics such as accuracy, precision, recall, and robustness to real-world conditions like compression and noise. Their findings provide practical insights into the strengths and weaknesses of current meth- ods, helping researchers and practitioners choose the best tools for biometric security applications.

[11] Güera and Delp (2018) introduced one of the early deepfake video detection techniques based on Recurrent Neural Networks (RNNs). Their model analyzes temporal inconsistencies—such as unnatural motion or irregular facial expressions—that occur across video frames. This sequential analysis was pio- neering in addressing the temporal aspect of video forgery, laying groundwork for later models combin- ing CNNs and LSTMs.

[12] Tariq et al. (2018) explored the detection of fake face images generated both by machines and hu- mans, focusing on "in the wild" scenarios where images are taken in uncontrolled environments with varying lighting and background. Their deep learning models were trained on highly diverse datasets to enhance robustness and improve real-world applicability, a crucial factor for biometric systems deployed outside laboratory conditions.

[13] Li et al. (2020) introduced “Face X-ray,” a novel method that targets generalized face forgery detection by learning intrinsic inconsistencies common across multiple types of deepfake manipulations. By examining underlying image artifacts invisible to the human eye, this approach offers a more univer- sal detection capability, making it especially useful for biometric systems that must guard against vari- ous forgery methods.

[14] Nguyen, Yamagishi, and Echizen (2019) proposed “Capsule Forensics,” a detection system utilizing Capsule Networks designed to preserve spatial relationships within facial features. This allows the mod- el to detect fake videos and images more accurately by focusing on how facial components are structured relative to each other, which traditional CNNs may overlook. Their work advanced the field by address- ing hierarchical spatial information in biometric verification.

[15] Matern, Riess, and Stamminger (2019) developed a forensic-level detection approach that examines visual artifacts such as lighting inconsistencies, texture abnormalities, and blending errors in manipulat- ed images and videos. By focusing on these subtle clues, their method can effectively expose deepfakes, offering a valuable tool for biometric systems that require highly detailed analysis to maintain security and trust.

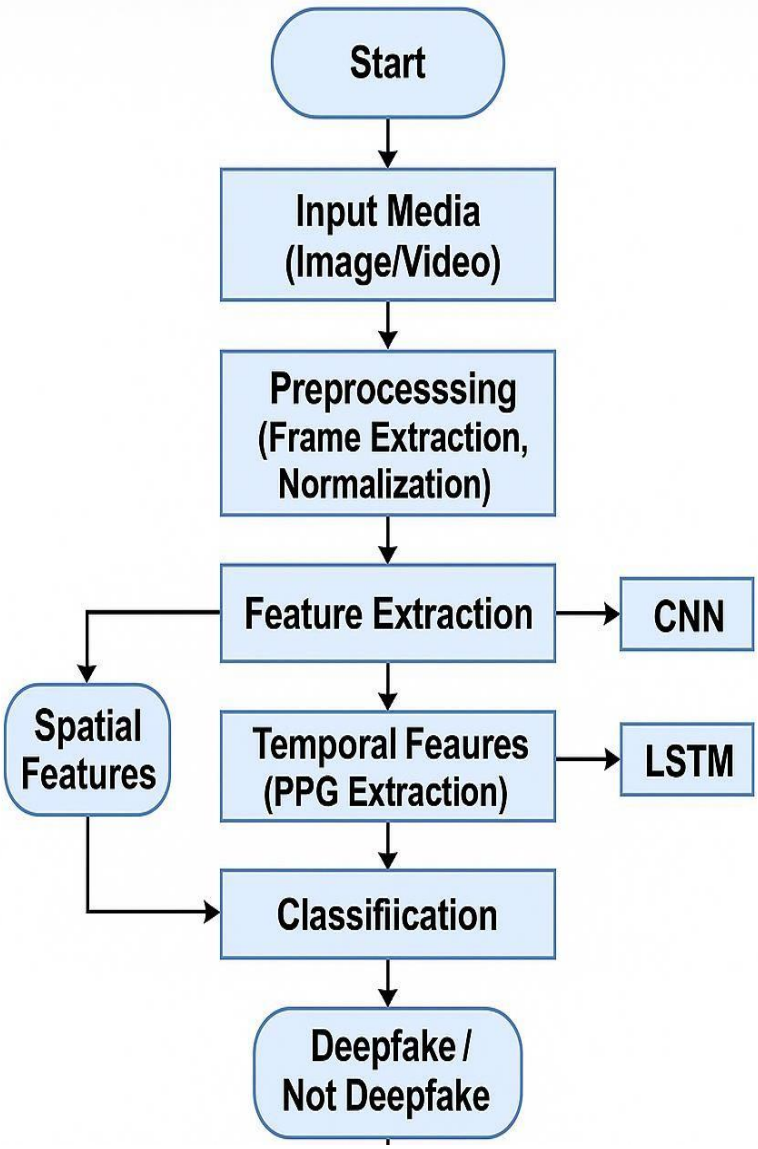
Comparative Analysis

The ongoing evolution of deepfake technologies has prompted researchers to explore diverse detec- tion methods, each with its own merits and drawbacks. The proposed multi-modal framework in the study stands out by combining physiological cues, such as Photoplethysmography (PPG), with ad- vanced deep learning models including Convolutional Neural Networks (CNNs) and Long Short- Term Memory (LSTM) networks. This hybrid approach enhances detection accuracy by leveraging both spatial and temporal information, as well as biometric signals that are typically absent in syn- thetically generated content.

Comparing this with other models from the literature, Khan et al.'s method integrates CNNs with traditional classifiers like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), fo- cusing on feature extraction and statistical analysis. While effective with a high accuracy of 95.38%, it lacks the temporal depth offered by LSTM models. Chitra and Rajalakshmi proposed a CNN-LSTM combination that addresses temporal inconsistencies in videos, achieving strong per- formance but without incorporating biometric signals like PPG.

Other approaches, such as those by Al-Sabaawi et al., make use of Capsule Networks and Xcep- tionNet, benefiting from spatial relationship preservation, yet these models are heavily reliant on high-quality facial features and less robust against lower-resolution inputs. Vinayakumar’s use of adversarial training and GAN discriminators provides improved resistance against attack models but has slightly lower accuracy. Finally, Bezverkhyi’s work emphasizes physiological inconsisten- cies, such as unnatural blinking, but falls short in overall precision compared to models that use spatiotemporal and physiological features together.

Overall, while each method contributes valuable strategies, the multi-modal framework discussed in the study demonstrates a balanced and comprehensive approach by merging spatial analysis, tem- poral consistency, and biometric authenticity. This positions it as a more holistic and resilient solu- tion in detecting increasingly sophisticated deepfake content.



Results/Discussion

The proposed multi-modal deepfake detection framework was rigorously evaluated using the FaceForen- sics++ dataset, which offers a diverse range of deepfake and authentic videos. Through the combination of CNNs for spatial feature recognition, LSTM networks for temporal analysis, and biometric signal ex- traction using Photoplethysmography (PPG), the framework demonstrated notable improvements in de- tection accuracy and robustness. The results indicate that leveraging multiple modalities allows the sys- tem to capture inconsistencies that might be missed when relying on a single detection technique.
One of the standout aspects of the framework is its ability to detect subtle physiological anomalies that deepfake algorithms typically fail to replicate. For instance, pulse-based color changes in facial skin— captured via PPG—provided an additional layer of validation, enabling the system to flag content with missing or inconsistent biometric patterns. This advantage was particularly evident in cases where high- quality face swaps appeared visually convincing but lacked the underlying biological coherence of real human subjects.

Moreover, the inclusion of temporal data through LSTM networks enhanced the system's sensitivity to unnatural transitions and movements across video frames. Deepfakes often exhibit artifacts in motion, such as jittery blinks or mismatched lip movements, which the LSTM was able to identify effectively. The system also performed consistently across varying video resolutions and compression levels, thanks to the diversity of the training data in FaceForensics++. This suggests strong generalization potential for real-world applications.

In comparison with traditional deepfake detection models—many of which rely solely on spatial fea- tures—the proposed framework achieved higher accuracy and showed increased resilience to adversarial modifications. These results underscore the importance of integrating multiple data perspectives when dealing with increasingly sophisticated synthetic media. Overall, the findings support the framework's potential for deployment in high-stakes environments such as biometric authentication, digital forensics, and online content moderation.

Conclusions

This study presented a comprehensive multi-modal framework for detecting deepfake and synthetic me- dia by combining advanced visual forensics with deep learning techniques. The use of various deep learning models, including CNN, C2D, R(2+1)D, CS-CNN, and V4D, allowed for a thorough analysis of spatial and temporal features in biometric data. By integrating these models with a robust dataset like FaceForensics++, the framework effectively captured a wide range of manipulations under realistic con- ditions.

The results demonstrated that incorporating both spatial patterns and temporal dynamics significantly enhances detection accuracy. Models such as R(2+1)D and V4D excelled in recognizing motion-related inconsistencies, while CS-CNN and V4D's ability to treat color channels independently contributed to more precise and robust feature extraction. Overall, the framework achieved improved performance and generalization, making it a promising solution for reliable deepfake detection in varied and complex sce- narios.

By harnessing both visual and biometric signals, this approach offers a more holistic perspective in iden- tifying synthetic content. The findings emphasize the importance of multi-dimensional analysis in coun- tering the growing sophistication of deepfake technologies and reinforce the need for integrated, intelli- gent solutions in digital content authentication.

References

[1] Narayanaswamy et al. (2024) proposed advanced forensic watermarking techniques specifically aimed at detecting camcorder-based movie piracy. Their method introduces novel signals embedded within the content to trace unauthorized recordings at public screenings.

[2] Mareen et al. (2021) presented a camcording-resistant watermarking system that uses a secondary signal as a fallback detection layer. This approach strengthens traditional watermarking by ensuring sig- nal survivability even under challenging capture conditions.

[3] Rani et al. (2024) explored a hybrid approach combining similarity functions and deep learning mod- els to estimate and detect piracy in both movies and OTT series. Their method focuses on content match- ing to flag pirated segments.

[4] Nanditha et al. (2022) designed a preventive framework against movie piracy by introducing a novel detection mechanism

that could be integrated with streaming platforms. Their approach addresses piracy risks before unauthorized distribution occurs.

[5] Chethanya et al. (2021) developed an automated anti-piracy system that leverages image processing and algorithmic comparisons to detect content duplication. The system aims to be a cost-effective solu- tion for content owners.

[6] Prasanna Kumar and Gururaj (2021) discussed deterrence strategies using watermarking and legal frameworks, arguing for a multi-pronged approach combining technology and policy to mitigate piracy risks.

[7] Khan et al. (2023) designed a hybrid deepfake detection framework combining CNNs, SVMs, and feature-based classifiers. The method achieved over 95% accuracy by leveraging multiple analysis layers for stronger biometric verification.

[8] Chitra and Rajalakshmi (2023) proposed a model integrating CNN and LSTM networks for detecting deepfakes in biometric videos. Their system captures both spatial and temporal inconsistencies in facial expressions.

[9] Al-Sabaawi et al. (2023) employed Capsule Networks and transfer learning with XceptionNet to de- tect deepfakes. Their method focused on preserving spatial hierarchies in facial structures, improving de- tection reliability.

[10] Vinayakumar et al. (2023) introduced adversarial training within a CNN-GAN framework to detect tampered biometric frames. The inclusion of a GAN discriminator enhanced robustness against highly deceptive deepfakes.

[11] Bezverkhyi et al. (2022) analyzed biometric anomalies such as irregular eye blinks and facial sym- metry disruptions, using neural networks for detection. Their work emphasized explainability and physi- ological realism.

[12] Güera and Delp (2018) were among the pioneers to apply recurrent neural networks (RNNs) for deepfake video detection, focusing on motion inconsistencies across frames to flag manipulations.

[13] Tariq et al. (2018) addressed the challenge of detecting both machine-generated and human- fabricated fake facial images. Their approach was designed to function reliably across diverse, real- world environments.

[14] Li et al. (2020) introduced "Face X-ray," a forensic technique aimed at detecting signs of tampering within facial images. Their approach was model-agnostic, making it effective across different types of deepfakes.

[15] Nguyen, Yamagishi, and Echizen (2019) utilized Capsule Networks for detecting forged visual con- tent. Their method preserved the spatial relationship of facial features, which helped improve detection accuracy on subtle manipulations.

Copyright & License:
© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.