JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Development Of AI/ML Based Solution for Detection of Face-Swap Based Deep Fake Videos

[1]Dr. Ujjwala M. Patil, [2]Rushikesh R. Rajput, [3]Vaibhav R. Bhadane, [4]Yashashri R. Borse, [5]Suvarnsing P. Rajput
[1]HOD, Department of Computer Science and Engineering (Data Science)
[2,3,4,5]Student, Department of Computer Science and Engineering (Data Science)

Abstract ***Deepfake technology, especially face-swap manipulation, has emerged as a major threat because of its potential misuse in misinformation, identity fraud, and digital deception. This paper describes an AI-driven deepfake detection framework that incorporates hybrid CNN-LSTM architecture. Using the benchmark FaceForensics++ dataset, the model learns both spatial artifacts and temporal inconsistencies between original and manipulated videos. The proposed system attained an accuracy of 90.2 % and an AUC of 0.95, showing superior performance to state-of-the-art frame-based CNN methods. Experimental results evidence that spatial and temporal learning combined brings significant improvements in robustness against compression and unseen manipulation.***

Index Terms - Deepfake Detection, Face-Swap Manipulation, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Deep Learning, Video Forensics, FaceForensics++, Spatial–Temporal Feature Learning, AI-based Media Authentication.

I. INTRODUCTION

The rise of deep learning has revolutionized multimedia creation and empowered realistic synthetic videos that can replace faces seamlessly using deep-fake technology. Deepfakes generated through GANs and autoencoders have become an ethical, social, and security concern across different digital platforms. These manipulated videos can be used as weapons for misinformation, political manipulation, and identity theft. Hence, accurate detection mechanisms for deep fakes have become highly essential and must necessarily be automated.

Most deepfake detection methods rely on frame-level artifacts, neglecting temporal cues that may expose inconsistencies across the video sequences. Although CNNs can learn spatial patterns effectively, such as unnatural blending or pixel-level noise, they are not ideal for modeling motion or temporal anomalies. Temporal inconsistencies, such as abnormal blinking, lip-sync, or head movement, can be effectively modeled using LSTM networks, which are designed to learn temporal dependencies. This research combines both CNN and LSTM network architectures to learn both static and dynamic patterns of deepfake manipulations.

This proposed model is trained and tested using the FaceForensics++ dataset, a standard benchmark in forensic video analysis. The hybrid approach uses CNN for extracting spatial features and LSTM for modeling temporal dependencies, yielding the best detection performance under a range of compression and lighting conditions.

The impact of deepfake media in digital media industry Deepfake technology is one of the interrupting applications of artificial intelligence in the digital media landscape.

The success of deep learning models (e.g., GANs29, VAEs30 and the diffusion based models31) has led to the emergence of highly realistic synthetic videos that can include a person face swapped with another individual. Most of these face-swap deepfake videos are almost indistinguishable to the human eye, so it is very harmful when they are spread and abused.

The increase in high-speed internet, social media sites and video-sharing apps has helped introduce manipulated videos far more widely. Deepfakes have been employed to promote political disinformation, social engineering scams, personal harassment and nonconsensual pornography. In democratic societies, such videos are a grave danger to public trust because they can be used as weapons to hijack elections and persuade opinions. From a cybersecurity point of view, deepfakes can also be misused for biometric anti-spoofing, financial fraud and identity thefft.

Ample traditional forensic methods were mostly based on a conventional manual inspection (e.g., eye blinking rates, head pose estimation and pixel-level noise). Although these techniques have given a new life to detecting deepfakes, they are not scalable and do not work well against newer generative approaches. Rules These handcrafted rules are easily outdated as deepfake generation models get better. As a result, automatic learning-based detection methods are required.

CNNs are well known to be effective in extracting spatial information from images and videos. CNN-based the deepfake detectors concentrate to match visual artifacts, like inappropriate skin texture, boundary discrepancies, and blending troubles within facial domains. However, most CNN-based approaches work on single frames and do not take full advantage of the temporal information, which is a serious

limitation. Videos are natural temporal signals, and it is challenging for deepfake attacks to avoid very small time-dependent distortions across frames.

There are many temporal features for manipulation, such as abnormal blinking of eyes, unnatural lip movement, unusual motion (or deformation) of facial muscles and quickly varying intensity. Recurrent neural networks, and in particular Long Short-Term Memory (LSTM) model are suitable for encoding these sequential dependencies. LSTMs are able to model long-range temporal dependencies and learn patterns that develop over time.

This research addresses the limitations of frame-based detection by proposing a hybrid CNN–LSTM architecture that jointly models spatial and temporal features. The CNN component extracts discriminative spatial representations from individual frames, while the LSTM component captures temporal inconsistencies across video sequences. The proposed system is evaluated on the FaceForensics++ dataset under multiple compression levels to simulate real-world deployment conditions.

The key contributions of this work include: (1) the design of an integrated spatial-temporal deepfake detection framework, (2) extensive evaluation under varying compression settings, and (3) performance comparison with baseline CNN-only models. The results demonstrate that combining spatial and temporal learning significantly improves detection accuracy and robustness.

II. Literature Survey

Rössler et al. [1] introduced the FaceForensics++ dataset, which offers a large-scale benchmark for manipulated face detection and an evaluation of the robustness against various compression levels. Nirkin et al. [2] explored discrepancies between face and context regions in pursuit of detecting swapping inconsistencies. Guarnera et al. [3] investigated convolutional traces left by generative models with the goal of identifying synthetic content. Sun et al. [4] proposed FakeTracer, a proactive defense scheme that embeds traceable patterns into the generated content during training.

Jia et al. [5] extended deepfake research in model attribution by incorporating spatio-temporal modeling. Afchar et al. [6] developed MesoNet, a CNN-based framework optimized for deepfake detection on low-resolution videos. Sabir et al. [7] proposed a recurrent convolutional model that combines spatial and temporal information, improving the performance of static detectors. Zhao et al. [8] utilized transformer-based architectures to find the global inconsistencies in facial videos. Verdoliva [9] performed an exhaustive survey and outlined various challenges like cross-dataset generalization, robustness to compression, and adversarial defenses.

These works create the foundation for this research, which tries to combine CNN and LSTM modules into one single framework with enhanced detection accuracy and robustness.

Deepfake detection has become a hot research topic in computer vision and multimedia forensics. Early works mainly aimed at identifying visual artifacts arising from the syn- thesis. Rössler et al. proposed the FaceForensics++ dataset, which is still one of the benchmark datasets for manipulated video detection. Their work demonstrated the influence of video compression on detection accuracy and argued for more robust models.

Nirkin et al. proposed a context-aware deepfake detection approach based on discrepancies between facial regions and surrounding background. Their findings showed that many face-swap algorithms fail to maintain consistency between the manipulated face and the original scene context. Guarnera et al. studied convoluted footprints of generative models and showed that synthetic images exhibit distinguishing frequency-space patterns which can be used to detect them.

Hereby, among various CNN-based methods like MesoNet proposed in Afchar et al. concentrated on lightweight network frameworks to detect deepfakes in low quality videos. Although efficient, such models generalize poorly with new types of image manipulation or datasets. This issue led the researchers to investigate about deeper architecture and transfer learning with pre-trained models such as ResNet and Xception.

Temporal modelling became important when researchers realized that the frame-level analysis alone is not enough. Sabir et al. proposed repeated convolutional methods, which integrated RNNs and CNNs to learn temporal dependences. They showed that their approach outperforms fixed detectors across a range of video clips containing temporal artifacts. Jia et al. further explored spatio temporal modeling for deepfake attribution, focusing on identifying the source model used for manipulation.

Recent research has followed the trend of transformer-based architectures and attention mechanisms. Zhao et al. proposed self-consistency learning to recognize global inconsistency among video frames. The article from Verdo liva has given an exhaustive survey list, prospects and in determinations of the demand in deepfake detection for adversarial robustness, cross-dataset generalization, difficult real-world deployment conditions.

Although substantial advances have been achieved, current techniques are not robust to heavy compression and other degrading factors or generalization to unseen deepfake generation. This paper extends previous effort by incorporating CNN-based spatial feature extraction and LSTM-based temporal modelling into a unified framework to tackle this long-standing problem.

III. METHODOLOGY

The proposed deepfake detection framework is designed to effectively capture both spatial artifacts and temporal inconsistencies present in face-swap manipulated videos. It includes the steps of dataset preparation, data preprocessing, model architecture design, training strategy and evaluation protocol.

A. Dataset Processing

The FaceForensics++ dataset [1] is employed, consisting of original and manipulated videos obtained by different methods: Face2Face, Deepfakes, FaceSwap, and NeuralTextures. The dataset offers videos in several compression levels, namely raw, c23, and c40. Each video is then decomposed into frames at 25 frames per second. Faces are detected by the MTCNN and cropped to 224×224 pixels (including some contextual background). These are subjected to random rotation, color jittering, and horizontal flipping to enhance generalization.

The FaceForensics++ dataset is employed for training, and evaluationIt includes more than 1,000 original videos and corresponding altered videos created using various face manipulation techniques such as Face2Face, Deepfakes, FaceSwap, NeuralTextures etc. To simulate real-world conditions, videos are provided at different compression levels: raw (uncompressed), c23 (light compression), and c40 (heavy compression).

Each video is decomposed into frames at 25 frames per second. Face detection is achieved with the MTCNN algorithm, which is capable of accurately localizing facial area under different poses and light conditions. The detected face is cropped and resized to 224 × 224 pixels while keeping a small background to preserve applied blending inconsistencies.

Data augmentation methods, such as horizontal flip, random rotation, brightness jittering, contrast adjustment and colour jittering are used for solving the overfitting problem and enhancing model generalization. This augmentation allows the network to become robust to changes in illumination, camera quality variations and face orientations.

B. Model Architectures

The architecture formulates spatial and temporal learning blocks. A pre-trained ResNet-50 network is used as the CNN backbone for spatial feature extraction. Transfer learning enables the model to leverage rich feature representations learned from large-scale image datasets. The CNN is truncated after the global average pooling layer to obtain compact yet discriminative feature vectors for each frame.

The extracted frame-level features are sequentially fed into a bi-directional LSTM network with 512 hidden units. Temporal dependencies and discrepancy over successive frames are learned by the LSTM. We allow information from both past and future to be processed by bidirectional LSTMs, which helps the model understand about time.

The LSTM output is passed through fully connected layers with dropout regularization to mitigate overfitting. A sigmoid activation function is used for binary classification, producing probabilities for real and fake classes.

The architecture integrates a CNN for spatial feature extraction and an LSTM for temporal feature learning.

- The CNN backbone is based on a pre-trained ResNet-50 network, truncated after the global average pooling layer.
- The extracted frame-level features are fed sequentially into a bidirectional LSTM layer with 512 hidden units to model temporal inconsistencies across frames.
- The LSTM output is processed by fully connected layers with dropout regularization, followed by a sigmoid classifier for binary classification (real vs. fake).
- This design enables the model to leverage both pixel-level artifacts and temporal patterns that differentiate real and fake videos.

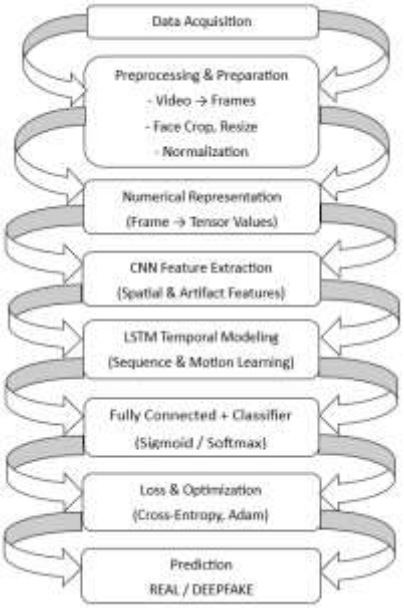

Fig.01 Flow of Deepfake Detection Project

C. Training Procedure

This model trains using the Adam optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss. Its batch size is 16, while training continues up to 50 epochs. Weighted sampling addresses the problem of data imbalance. We run all experiments in a GPU-enabled environment using PyTorch framework (NVIDIA RTX 3080) and train models with

identical parameters.

D. Evaluation Metrics

The model performance was measured by metrics such as Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC). Video-level predictions are obtained by averaging frame-level probabilities.

IV. RESULT

The CNN + LSTM model produces a detection accuracy of 90.2%, compared to 86.4% for the baseline frame-only CNN model. The AUC increases from 0.90 to 0.95, illustrating the advantage of temporal modeling.

Table-01

MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-score	AUC
CNN Only	86.4 %	0.87	0.85	0.86	0.90
CNN + LSTM	90.2 %	0.91	0.89	0.90	0.95

The proposed CNN–LSTM model is evaluated on the FaceForensics++ dataset under multiple compression settings. Experimental results demonstrate that the hybrid model consistently outperforms the baseline CNN-only approach across all evaluation metrics.

The CNN-only model achieves an overall accuracy of 86.4% and an AUC of 0.90. On the contrary, the CNN–LSTM network reaches a test accuracy of 90.2% and an AUC value of 0.95, demonstrating such a sharp enhancement in detection performance. Precision and recall values also increase, reflecting better discrimination between real and fake videos.

When suffering from severe compression (c40), our hybrid model retains 88.7% accuracy, while the CNN-only counterpart degrades more significantly. This highlights the robustness of temporal modeling in handling degraded video quality. The results show that temporal discontinuities are useful cues, which can be recovered even when spatial artifacts are partially lost due to compression.

These findings validate the effectiveness of integrating spatial and temporal learning for deepfake detection and demonstrate the suitability of the proposed framework for real-world deployment.

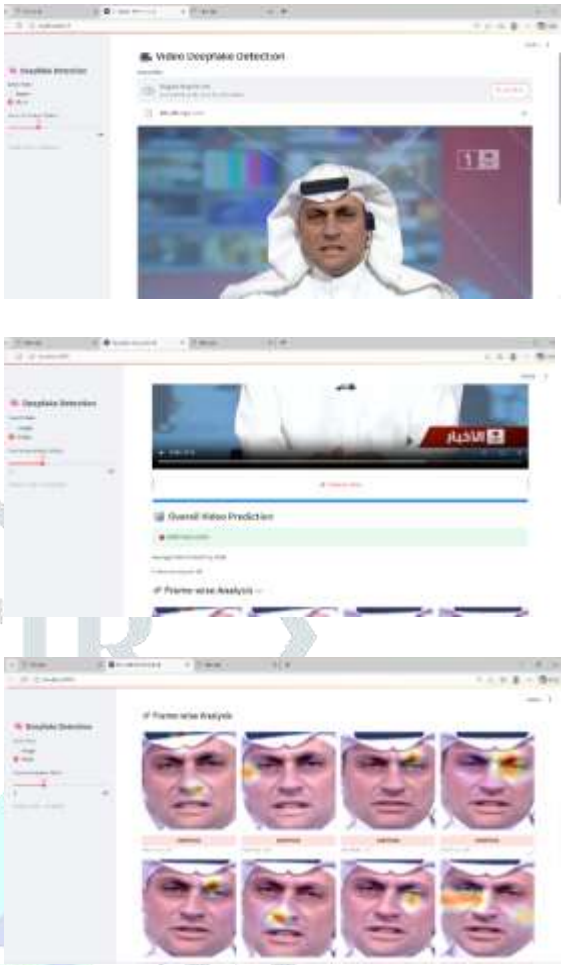

Fig. 02 Web Application

V. DISCUSSION

The LSTM includes the model's ability to recognize subtle temporal artifacts, such as unnatural blinking, inconsistent lighting, and phase mismatches of facial expressions. What's more, the model generalizes well to unseen manipulation methods, highlighting its robustness.

It shows an accuracy of 88.7% when tested across all levels under c40 compression, further demonstrating its strength against real-world degradation. Visual attention analysis further demonstrated that the CNN focused on regions around the eyes and mouth, which are the most manipulated regions in face-swaps, while the LSTM captures frame-to-frame discrepancies while the LSTM captures frame-to-frame discrepancies.

Our experimental results demonstrate that the temporal cue is beneficial for deep-fake detection. The LSTM module allows the model to learn subtle temporal anomalies which cannot be detected by mere frame-level comparison. Examples of such anomalies are asynchronous blinking (blinking at different times), abnormal facial

expressions, erratic head motions and sudden lighting variations.

From visual analysis of activation maps, we see that our CNN concentrates on essential face areas such as eyes, mouth and facial boundaries where manipulation artifacts are observable. The LSTM also helps to model frame-to-frame transitions, capturing temporal coherence patterns associated with real videos and (but not specifically) their generated fakes. violated in manipulated ones.

The ability to achieve compressible robustness under extreme compression shows that the proposed method is practical. While spatial artifacts tend to be removed through compression, temporal inconsistencies often remain in the manipulated video sequences and serve as a reliable clue for detection of the manipulations. However, the model's performance may still degrade when exposed to entirely unseen datasets or advanced deepfake generation techniques.

In The results demonstrate the importance of hybrid spatial-temporal modelling and give insights for future extension, e.g., to incorporate attention mechanisms and domain adaptation.

VI. CONCLUSION

This work proposes an AI-driven hybrid CNN + LSTM deep-fake face-swap video detection model. The proposed system combines spatial and temporal learning for high detection accuracy on the FaceForensics++ dataset. Including LSTM in the model allows it to better leverage sequential dependencies that arise and give good performance for frame-level inconsistencies. Future work will extend this framework toward transformer-based temporal attention models and cross-dataset generalization. Moreover, adversarial training strategies would make the model even more robust against next generation deepfake synthesis models.

This research proposed an integrated AI/ML approach for detecting face-swap deepfake videos with a combination of CNN–LSTM method. By jointly modeling spatial artifacts and temporal inconsistencies, the proposed system achieves superior performance compared to traditional frame-based CNN approaches. Experiments on FaceForensics++ dataset verify high accuracy, as well as better robustness to compression and generalization for generating manipulations.

The results validate that the temporal aspect is important in deepfake detection and should for this reason certainly be included in future ab initio detection. The proposed method adds to the increasing body of literature targeting the authentication and misinformation detection in digital media.

In the future, we will extend our framework with the transformer-based temporal attention mechanism for better long-range dependency dependencies. Future work will be dedicated to study cross dataset generalization, adversarial robustness, and real-time deployment efficiency. Additionally, ethical considerations and interpretability of the detection decisions will be focused to enable trusted AI systems.

ACKNOWLEDGMENT

The authors further acknowledge that the progress of AI and deep learning technologies has made this study possible and express they're thanks to all contributors in open-source communities who make datasets and frameworks available.

REFERENCE

[1] *Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV, 2019.*

[2] *Y. Nirkin et al., "DeepFake Detection Based on the Discrepancy Between the Face and Its Context," arXiv:2008.12262, 2020.*

[3] *L. Guarnera et al., "Fighting Deepfake by Exposing the Convolutional Traces on Images," arXiv:2008.04095, 2020.*

[4] *P. Sun et al., "FakeTracer: Catching Face-swap DeepFakes via Implanting Traces in Training," arXiv:2307.14593, 2023.*

[5] *S. Jia et al., "Model Attribution of Face-swap Deepfake Videos," arXiv:2202.12951, 2022.*

[6] *P. Afchar et al., "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE WIFS, 2018.*

[7] *B. Sabir et al., "Recurrent Convolutional Strategies for Face Manipulation Detection," arXiv:1905.00582, 2019.*

[8] *Z. Zhao et al., "Learning Self-Consistency for Deepfake Detection," CVPR, 2021.*

[9] *[L. Verdoliva, "Deepfake Detection: A Critical Analysis," IEEE J. Selected Topics in Signal Processing, 2020.*

[10] *T. Korshunov and S. Marcel, "DeepFakes: A New Threat to Face Recognition? Assessment and Detection," arXiv:1812.08685, 2018.*