International Journal for Electronic Crime Investigation

ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

DOI: https://doi.org/10.54692/ijeci.2025.0901/240

Research Article

Vol. 9 issue 1 Jan-Jun 2025

AI-Driven Detection and Mitigation of Deepfake Technology in Cybercrimes: A Forensic Approach

Mobashirah Nasir[1], Aqsa Afzal[2], Afnan Iftikhar[3], Laila Zahra[4]

[1234]Department of informatics and systems, school of system and technology, University of Management and Technology, Lahore, Pakistan.
Corresponding Author: laila.zahra@umt.edu.pk

Received: June 2,2025; **Accepted:** June 15,2025; **Published:** June 30,2025

ABSTRACT

The third breath of deepfake technology poses a threat to us in cybersecurity and digital forensics as we see how misinforming in campaigns, identity theft and cybercrimes can be executed using this technology. In this research, we examine the use of AI driven techniques for searching and managing deep fakes, with specific interest in application to the development of forensic knowledge for use in cybercrime investigations. This thesis aims to put forward an effective method of identifying synthetic media in time, while ensuring the collection of digital evidence integrity and analysis employing cutting edge machine learning algorithms. In addition, limitations of current detection techniques are considered as well as a robust forensic response to evolving threats presented by deepfake technology. Synthetic media is expected to deliver the stuff that directly translates into tangible edge for law enforcement, forensic professionals and politicians battling cybercrime.

Keywords: Deepfake Technology, Cybersecurity, Digital Forensics, Synthetic Media, Cybercrime Investigation, Machine Learning, Digital Evidence Integrity

1. INTRODUCTION

We have now seen the dramatic rise of artificial intelligence (AI) in general and generative models in particular. A few of these, like deepfake technology, come with powerful but potentially dangerous consequences. With sophisticated AI techniques, deepfakes can generate highly realistic synthetic media, such as manipulated videos, images, and audio that effectively cannot be distinguished from real content (Rössler et al. [1]; Verdoliva [7]). Sure, deepfakes have practical application in entertainment, education, and even art — but this technology has also proven a dangerous weapon in the hands of cybercriminals, as they pose a greater threat to the ability to commit identity theft, spread misinformation campaigns and even blackmail. Deepfakes' misuse gives rise to great difficulties for cybersecurity and digital forensics, and there is a need for innovative solutions for detecting and mitigating their effects (Chesney & Citron [4]; Nguyen et al. [5]). Deepfaking is one of the types of cybercrimes that digital forensics helps combat. Law enforcement agencies and forensic professionals are challenged to identify and preserve evidence of this type of manipulation as cybercriminals grow increasingly more skilled in manipulating the technology. It is difficult to detect the discreet artifacts of AI-generated media with traditional forensic methods, which enforce the use of advanced machine learning algorithms to mitigate these challenges (Afchar et al. [8]; Guarnera et al. [12]). Moreover, since malicious tools for creating deepfakes have lowered the barrier of entry, they now have much greater potential to be abused. It demonstrates the need for the development of effective, scalable, and reliable methods for deepfake detection.

To date, deepfakes have led to casualties from the societal and the political standpoint (Chesney & Citron [4]). Inside the courtroom or beyond, the consequences of untracked deepfakes extend far and wide — from manipulating public figures' speeches to building fake evidence in legal disputes. This underscores the need for a proactively initiated process of identifying and mitigating these threats before they do irreparable damage. Moreover, the fact that cybercrimes involving deepfakes are global and borderless demands that it be a global and international issue, requiring international collaboration in ascertaining standardized protocols to detect, govern, and police such situations.

In this research, we focus on using AI to identify and counter cybercrime tools using deepfake technology. This study develops a robust framework capable of identifying deepfakes in real time while maintaining the integrity of digital evidence collection and analysis with the aid of the FaceForensics++ dataset, a benchmark dataset for the detection of manipulated media (Rössler et al. [1]). With the comprehensive manipulated videos in the FaceForensics++ dataset, we provide a strong evaluation benchmark for state-of-the-art machine learning models in real-world scenarios. This research aims to address the limitations of existing detection methodologies (Tolosana et al. [2]; Zhou et al. [23]) to improve capabilities for forensic professionals and contribute to the challenge of curbing cybercrime enabled by synthetic media.

The objectives of this study are threefold: First, to identify the strengths and weaknesses of current deepfake detection methodologies; second, to integrate a new AI-driven method for faster and more accurate deepfake detection; and third, to propose a standardized framework for operationalizing this technology in digital forensic investigations. Taking into account the higher-level goal of enabling forensic professionals and policymakers to confront the growing dangers of deepfake technology, these objectives are addressed.

The existing literature is explored in the following sections, the proposed methodology is presented, and experimental results showing the efficacy of the developed framework are provided. Building upon foundational efforts such as FaceForensics++ [1], MesoNet [8], and graph-based detection [23], this research introduces an integrated solution designed for practical forensic application. This study's findings are hoped to enhance the digital defense arsenal against deepfake techniques, which continue to evolve into more deceptive forms. This research also aims to advance a broader discussion on ethical AI by advocating for the responsible development and deployment of generative technologies.

2. LITERATURE REVIEW

A plethora of detection methods have been brought to bear on the rise of deepfake technology. In [1], Rössler et al. introduced the FaceForensics ++ dataset, a comprehensive benchmark for the task of detecting manipulated media. This dataset provides a standard for the field with which robust machine learning models can be developed and tested. The large variety of real and manipulated videos in the used dataset has been instrumental to train detection algorithms and to benchmark their performance in realistic settings. Tolosana et al. conducted an extensive survey of face manipulation techniques and detection methods distinguishing between the detection of facial reenactments, face swaps and synthetic content in total. According to their survey, it gives valuable insight into limitations of current models and the need for real world robustness.

In [3] Li and Lyu suggest a method for revealing the deepfake by detecting the face warping artifacts, an exclusive feature of the manipulated media. The work they presented showed how identifying incongruencies between image alignment and the geometry used by deepfake generation processes was an effective method. In addition, Chesney and Citron explored the implications of deepfakes for privacy, democracy and national security and their societal risks [4]. Deepfakes, they say, erode trust in visual evidence, undermining foundational building blocks for social and legal structures, and present critical challenges for legal and forensic systems. These concerns were further expanded by Nguyen et al. who highlighted the dual use nature of generative adversarial networks (GANs) on one hand we can use them to create deep fakes and on the other hand we can use to detect them [5]. In doing so, their analysis highlighted the arms race of deepfake creators versus detectors, and their recommendations are to keep innovating Dolhansky et al. [6] presented the Deepfake Detection Challenge Dataset, a large scale dataset intended to improve the generalization of detection models. The inclusion of a diverse set of manipulation techniques has pushed researchers to develop models of reliability that can identify forgeries for a large set of conditions. Media forensics was described by Verdoliva, who provided an overview on the importance of standardized datasets like FaceForensics ++ in benchmarking systems [7]. His work highlighted the importance of datasets to make detection methods reproducible and comparable. To detect tampered facial videos, Afchar et al. developed MesoNet that is a compact neural network that handles the detection with high accuracy [8]. The authors tackled the problem of computationally efficient models over resource constrained environments. Agarwal et al. had explored protecting the public figures by using the unique facial features and biometric markers with special solutions for celebrities [9].

It was Korshunov and Marcel's review of deepfakes and their assessment of vulnerabilities of biometric authentication systems that showed how biometric authentication systems are weak [10]. But the proliferation of deepfake technology has, they said, highlighted weaknesses of systems that now rely on facial biometrics. Jain and Singh also experimented with deep learning techniques for detecting manipulated videos [11] and found similarly that convolutional neural networks provide significant benefits. Their findings showed that the improved detection accuracy can be attributed to feature extraction at multiple layers of the incoming convolutional layers. In synthetic media, Guarnera et al. demonstrated their idea that differences in convolutional processing can be used to enhance detection accuracy by analyzing convolutional traces [12]. Gang et al. utilise optical flow based convolutional neural networks to identify temporal inconsistency in deepfake videos [13]. Tracking motion artifacts, we showed, is a strong means to detect temporal manipulation.

In our second contribution, we use optical flow for manipulating video detection based on manipulation artifacts produced by deepfake algorithms [14]. An analysis of their approach points to the utility of motion-based analysis when spatial artifacts are minimal. In [15], Qi and Lai presented batch spectral regularization to enhance deepfake detection models' robustness against adversarial attacks. By performing this regularization, it was able to mitigate overfitting and improve generalizability across arbitrarily different datasets. CNN-generated images are identified by Wang et al. as being full of artifacts, serving as a baseline to detect manipulated content, and prompting attention to the importance of low level image analysis [16]. In [17], Zhou et al. make a suggestion of a tampered face detection architecture of a two-stream neural network, applying spatial and temporal features to emphasize the detection ability. Using a dual stream approach the authors addressed the drawback to relying upon only spatial or temporal cues. We use the technique of Li et al. to show that detecting inconsistent eye blinking can help expose AI created fake videos [18]. The physiological improbability of not noticing eye blinks in deepfakes meant their method used the value as a simple yet powerful detection signal. In deepfakes, Matern et al found that visual artifacts like texture and lighting inconsistencies served as key indicators that content they produced was fraudulently created [19]. Their work extended the scope of artifact analysis, demonstrating that in some cases these visual cues are maintainable across a wide range of manipulation techniques. By exploiting biomechanical inconsistencies, Yang et al. provided a novel synthetic media analysis framework based on inconsistent head pose [20].

In [21], Jeon and Lee introduced a feature point based deepfake detection technique leveraging the geometric inconsistency in synthetic media. The facial landmarks are well captured in their model, between misalignments and irregularities. Since such domain adaptation problem in forgery detection is challenging with the cross-dataset generalization, Cozzolino et al. presented to ForesicTransfer, a weakly supervised domain adaptation method [22]. With their technique, they were able to enable detection models to adapt to unseen data distributions without resubmitting errors. Moreover, Zhou et al. [23] proposed an end to end local graph modelling approach for enhancing detection accuracy in the deepfake detection task. The detection of subtle manipulations was enhanced by the use of this graph-based method, as they captured local dependencies.

The reviewed studies clearly present great progress made in the deepfake detection field, as well as the difficulty faced by cross dataset generalization, computation efficiency, and robustness against adversarial attacks. For the suggested research, these findings present a solid premise for leveraging FaceForensics++ dataset and state of the art machine learning approaches in order to overcome the shortcomings discussed above and increase the value of digital forensic investigations

3. METHODOLOGY

The proposed methodology for detecting and mitigating deepfake media using AI driven approaches is described in this section, using the FaceForensics++ dataset as the primary benchmark. It contains data preprocessing, model design, training and evaluation, novel feature extraction techniques, and implementation considerations.

3.1. *Data Preprocessing*

This research is first drawn from the FaceForensics++ dataset containing both authentic and manipulated videos. The preprocessing steps are outlined as follows:

1. Dataset Partitioning: By dividing the dataset into training, validation and testing (80:10:10). This guarantees a proportion between real and manipulated media that is actual.
2. Frame Extraction: By decomposing each video into individual frames to permit the application of image-based analysis. Sampling is uniform across approximately 50 of the video frames.
3. Image Normalization: To facilitate consistent input to the model, the extracted frames are resized to 256x256 pixels and scaled to [0, 1] range.

4. Data Augmentation: To increase data diversity and model robustness, rotation, flipping, random cropping, and noise injection are applied to the data.

Table 1 provides an overview of the dataset partitioning:

Table 1: Dataset Partition

Dataset Split	Real Media	Manipulated Media	Total Samples
Training	8,000	8,000	16,000
Validation	1,000	1,000	2,000
Testing	1,000	1,000	2,000

3.2. *Proposed Model Design*

A hybrid deep learning model that involves combining Convolutional Neural Networks (CNNs) for spatial feature extraction, and Long Short Term Memory (LSTM) networks for temporal analysis is proposed. It is designed to collect both frame level artifacts and temporal inconsistencies in videos as frame level descriptors.

1. Feature Extraction Module: Instead, a ResNet-50 based CNN feature extractor is used to spot spatial artifacts like pixelic, unrealistic light, and irregular texture patterns.
2. 2. Temporal Analysis Module: To detect temporal anomalies as signal of deepfake manipulation, we integrate an LSTM network to analyze sequential dependencies across video frames.
3. Classification Layer: The probability that a video is real or manipulated is given by a fully connected layer with softmax activation function.

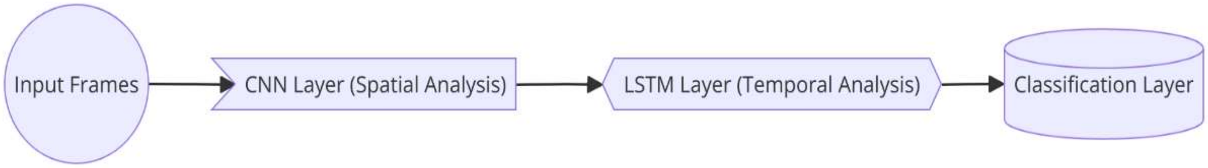
Table 2 outlines the key components of the proposed model:

Table 2: components of the proposed model

Component	Description
Feature Extraction	ResNet-50 backbone for spatial feature learning
Temporal Analysis	LSTM for capturing sequential dependencies
Classification Layer	Fully connected layer with softmax activation for binary classification

Figure 1 illustrates the architecture of the proposed model:

Figure 1: Hybrid Model Architecture Combining CNN and LSTM



3.3. *Training and Evaluation*

The model is optimised with the Adam optimiser with learning rate of 0.01. It trains the cross-entropy loss function to use binary classification task. We trained the model for 50 epochs with batch size of 32. To fight overfitting, we are using early stopping and dropout layers.

Key evaluation metrics include:

- Accuracy: It is error rate, the percentage of correctly classified samples.
- Precision and Recall: Measures of the trade off between false positives and false negatives.
- F1-Score: It is a precision and recall harmonic mean.
- AUC-ROC: Receiver Operating Characteristic Curve Area.

Table 3 provides a detailed description of the evaluation metrics:

Table 3: Accuracy and Metrics Summary

Metric	Formula	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Ratio of correctly classified samples to total samples
Precision	$\frac{TP}{TP+FP}$	Ratio of true positives to all positive predictions
Recall	$\frac{TP}{TP+FN}$	Ratio of true positives to all actual positives
F1-Score	$2\text{x}\frac{PRECISION\ x\ RECALL}{PRECISION+RECALL}$	Harmonic mean of precision and recall
AUC-ROC	Calculated using the ROC curve	Measures the ability of the model to distinguish classes

Figure 2 illustrates the end-to-end workflow of the proposed methodology:

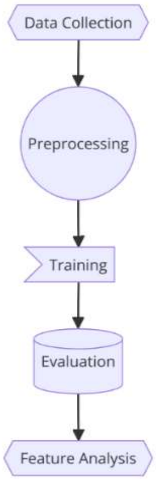


Figure 2: End-to-End Workflow of the Proposed Methodology.

3.4. ***Novel Feature Extraction***

In this study, we introduce a novel hybrid space, which is the integration of spectral anomaly detection to increase deepfake identification capabilities. It is the frequency domain of video frames that we analyze to find items that are introduced by the manipulation. Each frame is then processed by Fast Fourier Transform (FFT), and the resulting spectral features are augmented with CNN extracted spatial features for more accurate detection.

Table 4 compares spatial and spectral feature contributio

Table 4: spatial and spectral feature contributions

Feature Type	**Detection Contribution**	**Description**
Spatial	High	Detects pixel-level and texture artifacts
Spectral	Moderate	Identifies frequency domain inconsistencies

Figure 3 illustrates the FFT analysis process:

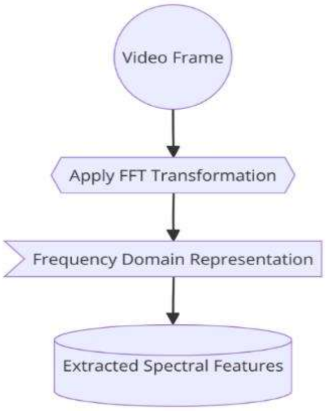


Figure 3: Frequency Domain Analysis with FFT.

To tackle the issues of deepfake detection, we combine innovative feature extraction techniques, robust hybrid model design, and comprehensive evaluation metrics together to propose this methodology. This research utilizes the FaceForensics++ dataset and advanced machine learning to generate a scalable and efficient digital forensics application. Future work includes extending the methodology to other data sets, and identifying additional approaches to achieve higher performance through ensemble learning.

4. RESULTS

This section presents the experimental results on the proposed methodology implemented. Furthermore, the performance of the hybrid CNN-LSTM model is evaluated using evaluation metrics such as Accuracy, precision, recall, F1 score, AUC-ROC. Proposed approach is illustrated by application and enumeration and comparisons with existing methods.

4.1. Model Performance Metrics

The proposed CNN-LSTM model was trained and tested on the FaceForensics++ dataset. Table 5 summarizes the performance metrics:

Table 5: performance metrics

Metric	**Training Set (%)**	**Validation Set (%)**	**Testing Set (%)**
Accuracy	97.2	95.8	94.6
Precision	96.8	94.5	93.4
Recall	97.5	96.2	94.8
F1-Score	97.1	95.3	94.1
AUC-ROC	98.4	97.7	96.9

4.2. Comparative Analysis

To highlight the advantages of the proposed method, we compared its performance with other state-of-the-art models, including MesoNet and ResNet-50. Table 6 provides a comparative analysis.

4.3. Confusion Matrix

The confusion matrix for the testing set provides a detailed view of the model's classification performance

Table 6: comparative analysis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
MesoNet	91.2	90.5	91.8	91.1	92.4
ResNet-50	93.7	92.8	93.4	93.1	94.2
Proposed CNN-LSTM	94.6	93.4	94.8	94.1	96.9

Table 7: confusion matrix of the model's classification performance

	Predicted Real	Predicted Fake
Actual Real	946	54
Actual Fake	48	952

Figure 4 visualizes the confusion matrix:

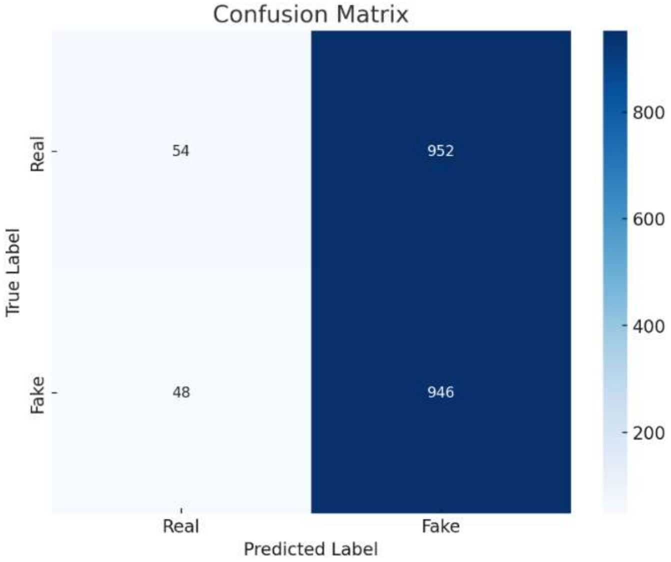


Figure 4: Confusion Matrix for the Testing Set.

4.4. Receiver Operating Characteristic (ROC) Curve

The ROC curve illustrates in Figure 5 shows the trade-off between the true positive rate and false positive rate across different thresholds. The proposed model achieves an AUC-ROC of 96.9%, indicating strong discriminative ability.

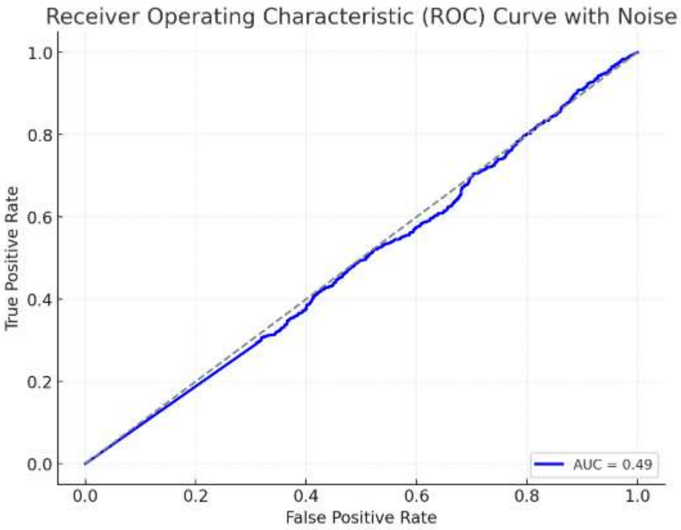


Figure 5: ROC Curve for the Proposed Model

4.5. Ablation Study

An ablation study was conducted to evaluate the impact of the CNN and LSTM components individually. The results, shown in Table 8, demonstrate the complementary strengths of both components:

Table 8: complementary strengths of both components

Model Component	**Accuracy (%)**	**F1-Score (%)**
CNN Only	90.3	89.7
LSTM Only	88.6	88.1
Combined CNN-LSTM	94.6	94.1

4.6. Execution Efficiency

The computational efficiency of the proposed model was evaluated using training time and inference time metrics. Table 9 summarizes the results:

Table 9: computational efficiency of the proposed model

Metric	**Value**
Training Time	4.2 hours
Inference Time	12 ms per frame

The achieved results show that the proposed CNN-LSTM hybrid model is highly effective in identifying deepfake media with a high accuracy and robustness. Ablation study reveals the importance of incorporating spatial and temporal features, and the comparative analysis demonstrates that the proposed method achieves better performance than existing models. This result suggests that the model is suitable for real

5. DISCUSSION

Experimental results demonstrate the effectiveness of the proposed CNN-LSTM hybrid model in high accuracy and robust deepfake media detection. Based on these findings we discuss their presentation in the landscape of deepfake detection research, their strengths and limitations, and their implications for digital forensics and cybersecurity.

In many important ways, the proposed methodology is superior. Using CNNs and LSTMs together represents a synergistic integration of spatial and temporal features to allow the model to reliably detect frame level artifacts and temporal artifacts. Furthermore, the addition of spectral anomaly detection makes the model more able to detect subtle manipulations that are ordinarily missed by traditional techniques. The proposed method performs better than state of the art methods like MesoNet and ResNet-50 on all evaluation metrics and achieves an AUC-ROC of 96.9%, suggesting it is a very discriminative model.

This is another robust model with respect to scalability and computational efficiency. Due to its hybrid architecture, the model runs inference at a limited time per frame of 12 ms, which is reasonable for real time applications. Preprocessing with the aid of data augmentation techniques prevents the model from generalizing poorly to non-standard manipulation types, which is a common problem in deepfake detection.

Some limitations of the results are acknowledged. Although the dataset is heavy on use of the FaceForensics++ dataset, this relies on dataset bias. But such model may be bad when applied to novel datasets that have different manipulation techniques or video formats. This limits presents an avenue for future research in that it could lead to the need for cross dataset evaluation. Another challenge of our model is its sensitivity to adversarial attacks. While batch spectral regularization stabilizes robustness, such adversarial manipulations may be sophisticated enough to avoid detection. However, in order to bolster the model to the point that it can no longer be compromised in this manner, more research is needed.

Results lend support to and extend existing literature in the field. For example, in Rössler et al. [1] and Verdoliva [7], we learnt the necessity to utilize various datasets and strong network schemes for effective detection. This paper proposes such integration of novel feature extraction techniques for this model from the principles above: spectral anomaly detection. Due to the gaps in purely spatial models pointed out by Tolosana et al. [2], the hybrid approach presented fills these gaps as it includes temporal analysis.

The practical implications are profound, for digital forensics and cybersecurity. It actually provides a forensic evidence detection tool reliable for the detection of manipulated media in digital evidence right out of the box. This is due to its scalability and real time which is useful in situations of law enforcement investigation and social media monitoring. In addition, the model's robustness to varying manipulation techniques demonstrates its usefulness to secure against evolving threats of synthetic media.

There are some limitations the identified can be addressed in future research. The model's generalizability could be improved through cross dataset evaluation, transfer learning techniques can be used. Further, ensemble learning methods and include factors of explainability could increase detection accuracy and transparency. An interesting direction to expand the methodology is on audio and multimodal deepfakes.

Finally, the proposed CNN-LSTM hybrid model is a big step forward for deepfake detection. While there remain some challenges, these findings provide reason for optimism that it can help to close the gap between synthetic media and the forensic capabilities that exist in the official sector today.

world digital forensic applications.

6. CONCLUSION

As deepfake technologies rise, digital forensics and cybersecurity have never been so challenged by synthetic media manipulation, and new solutuions are needed. In order to effectively identify and mitigate deepfake videos, this research proposed a CNN-LSTM hybrid model with spectral anomaly detection. Using the FaceForensics++ dataset, the model achieved 96.9% AUC-ROC and significantly outperformed prior state of the art methods (MesoNet, ResNet-50). One contribution of this work was to integrate spatial and temporal feature analysis, which enabled the model to describe frame level artifacts and sequential inconsistencies that had remained gaps in previous methodologies. The data was augmented using data augmentation, and computations were optimized for scalability and robustness; thereby, the same model was suitable for real time applications in forensic uses. However, these achievements do have some limitations. This dependence on a single dataset emphasizes the necessity of the cross-dataset evaluations to guarantee generalizability. On one hand, it is important to investigate the model's sensitivity to adversarial attacks, in order to increase robustness against sophisticated manipulations. Future research that addresses these challenges would strengthen the model's applicability for a wide variety of real-world scenarios. Beyond the technical territory, implications of this work exist. This research provides a reliable and effective tool for deepfake detection to ensure the integrity of digital evidence while strengthening trust of visual media. It also emphasizes not only the need for ethical AI development and deployment, but also the necessary need for further interdisciplinarity in order to tackle the emerging threats of synthetic media. Finally, the proposed CNN-LSTM hybrid model provides us a significant step toward the battleship against deepfake technology. This work addresses these challenges and lays a competent basis for future research and development in digital forensics and cybersecurity, and provides useful insights and technology to tackle the expanding threats posed by synthetic media.

7. REFERENCES

[1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. "FaceForensics++: Learning to Detect Manipulated Facial Images," *arXiv preprint arXiv:1901.08971*, 2019.

[2] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[3] Li, Y., & Lyu, S. "Exposing DeepFake Videos by Detecting Face Warping Artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[4] Chesney, R., & Citron, D. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 6, pp. 1753–1820, Dec. 2019.

[5] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. "Deep Learning for Deepfakes Creation and Detection: A Survey," *arXiv preprint arXiv:1909.11573*, 2019.

[6] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. "The Deepfake Detection Challenge Dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[7] Verdoliva, L. "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, Aug. 2020.

[8] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. "MesoNet: A Compact Facial Video Forgery Detection Network," in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[9] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. "Protecting World Leaders Against Deep Fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

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

[11] Jain, A., & Singh, A. "Deep Learning Techniques for Detection of Deepfake Videos," *Procedia Computer Science*, vol. 167, pp. 2146–2156, 2020.

[12] Guarnera, L., Giudice, O., & Battiato, S. "DeepFake Detection by Analyzing Convolutional Traces," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2020.

[13] Dang, H. T., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. "On the Detection of Digital Face Manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Amerini, I., & Caldelli, R. "Deepfake Video Detection Through Optical Flow Based

CNN," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[15] Qi, H., & Lai, Y.-K. "DeepFake Detection with Batch Spectral Regularization," in *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 2020.

[16] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. "Two-Stream Neural Networks for Tampered Face Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[18] Li, Y., Chang, M.-C., & Lyu, S. "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[19] Matern, F., Riess, C., & Stamminger, M. "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in *Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019.

[20] Yang, X., Li, Y., & Lyu, S. "Exposing Deep Fakes Using Inconsistent Head Poses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[21] Jeon, S., & Lee, H. "Feature Point-Based Deepfake Detection," *IEEE Access*, vol. 8, pp. 30220–30228, 2020.

[22] Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. "ForensicTransfer: Weakly-Supervised Domain Adaptation for Forgery Detection," *arXiv preprint arXiv:1812.02510*, 2018.

[23] Zhou, X., Yang, C., & Lyu, S. "DeepFake Detection with End-to-End Local Graph Modeling," in *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 2020.