Enhancing Security With AI-Media Fabrication Detection Through Advanced Neural Networks

[1]K. Manjula,[2]Bonthi Balanjali,[3]Rakshitha,[4]Thirupathi Sruthi,[5]Chinthala Sadhvika,[6]Ukkuturi Amulya

[1]Assistant Professor, Department of Computer Science & Engineering (AI & ML), Princeton Institute of Engineering & Technology For Women

[2,3,4,5,6]B. Tech Students, Department of Computer Science & Engineering (AI & ML), Princeton Institute of Engineering & Technology For Women

ABSTRACT

The rapid advancement of artificial intelligence has led to the creation of highly realistic fabricated media, commonly known as deepfakes. These manipulated images, videos, and audio files pose serious threats to digital security, privacy, and trust. This project proposes an advanced neural network-based system for detecting AI-generated media fabrications. The system utilizes deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to identify subtle inconsistencies and artifacts in media content. By analyzing spatial and temporal features, the model can accurately distinguish between authentic and manipulated media. The proposed system incorporates preprocessing, feature extraction, classification, and alert mechanisms to ensure efficient detection. It also supports real-time analysis and provides high accuracy compared to traditional methods. This approach enhances cybersecurity by preventing misinformation, identity theft, and digital fraud. The system is scalable, reliable, and capable of adapting to evolving fabrication techniques. Overall, the project contributes to strengthening digital media integrity and building trust in online platforms through intelligent AI-based detection mechanisms.

Keywords: Artificial Intelligence (AI), Deepfake Detection, Media Fabrication Detection, Advanced Neural Networks, Deep Learning, Convolutional Neural Networks (CNN), Multimedia Forensics, Generative Adversarial Networks (GAN) Detection, Cybersecurity, Synthetic Media Analysis, Digital Content Authentication, Image Forensics, Video Forensics, Misinformation Detection, Pattern Recognition.

I. INTRODUCTION

With the increasing use of digital platforms, the authenticity of multimedia content has become a major concern. Artificial intelligence has enabled the creation of highly convincing fake media, known as deepfakes, which can manipulate images, videos, and audio. These fabricated contents are often used for malicious purposes such as spreading misinformation, impersonation, and cybercrime. Traditional methods of detecting fake media are no longer effective due to the sophistication of modern AI techniques. Therefore, there is a need for advanced solutions that can accurately detect manipulated content. This project introduces a deep learning-based approach using advanced neural networks to identify AI-generated media fabrications. By leveraging CNNs for spatial feature extraction and RNNs for temporal analysis, the system can detect inconsistencies that are not visible to the human eye. The integration of machine learning models improves detection accuracy and efficiency. Additionally, the system provides real-time analysis and alerts to enhance security. This research focuses on developing a reliable and scalable solution to combat the growing threat of deepfakes and ensure the authenticity of digital media.

II. LITERATURE SURVEY

1. DeepFake Detection Using Deep Learning Methods

Authors: Y. Li and S. Lyu

Abstract:
This study investigates the detection of deepfake media using deep learning techniques. The authors propose algorithms that analyze inconsistencies in facial movements and visual artifacts generated by deepfake systems. By applying convolutional neural

networks to identify spatial and temporal anomalies, the approach achieves improved accuracy in identifying manipulated media. The research demonstrates that AI-based detection frameworks can effectively mitigate threats posed by synthetic media and enhance digital content verification.

2. FaceForensics++: Learning to Detect Manipulated Facial Images

Authors: A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner

Abstract:
This work introduces the FaceForensics++ dataset, a large-scale dataset designed for training and evaluating deepfake detection models. The authors evaluate several deep learning architectures including CNN-based models for detecting manipulated facial content in videos. Experimental results indicate that neural networks trained on large datasets can detect facial manipulations with high reliability, highlighting the importance of robust training data in AI-driven forensic systems.

3. MesoNet: A Compact Facial Video Forgery Detection Network

Authors: D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen

Abstract:
The MesoNet architecture focuses on detecting forged facial videos by analyzing mesoscopic image properties. The authors design lightweight CNN models capable of detecting deepfake videos even when computational resources are limited. The model achieves competitive performance while maintaining efficiency, making it suitable for real-time media verification applications.

4. Detecting Deepfake Videos Using Recurrent Neural Networks

Authors: S. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan

Abstract:
This research explores the use of recurrent neural networks combined with convolutional neural networks for detecting manipulated videos. The framework captures temporal inconsistencies across video frames that occur due to synthetic generation techniques. Results show that combining spatial and temporal feature extraction significantly improves detection performance.

5. Capsule Networks for Image Forgery Detection

Authors: T. Nguyen, F. Fang, J. Yamagishi, and I. Echizen

Abstract:
The authors propose a capsule network architecture for detecting image forgeries created using GAN-based approaches. Capsule networks capture hierarchical relationships between features, allowing the model to identify subtle inconsistencies in manipulated images. Experimental evaluation demonstrates improved generalization across multiple deepfake generation methods.

6. Exposing Deepfake Videos by Detecting Face Warping Artifacts

Authors: Y. Li and S. Lyu

Abstract:
This work presents a deepfake detection approach based on identifying artifacts caused by face warping processes during video synthesis. The proposed CNN-based method analyzes spatial distortions introduced during face replacement operations. The technique successfully detects manipulated videos and demonstrates the effectiveness of artifact-based detection.

7. Deepfake Detection Using EfficientNet Convolutional Networks

Authors: I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo

Abstract:
This research evaluates EfficientNet-based convolutional neural networks for deepfake detection tasks. The proposed model extracts discriminative features from manipulated media and achieves high classification accuracy. The authors emphasize the importance of transfer learning and large datasets for improving AI-based forensic analysis.

8. Multi-Task Learning for Deepfake Detection

Authors: H. Nguyen, J. Yamagishi, and I. Echizen

Abstract:
The study introduces a multi-task learning framework that simultaneously performs forgery detection and facial manipulation localization. By combining classification and segmentation tasks, the model improves detection performance and provides detailed explanations of manipulated regions. The results highlight the benefits of multi-task neural architectures in digital media forensics.

9. Detecting GAN-Generated Fake Images Using Deep Neural Networks

Authors: S. Tariq, S. Lee, H. Kim, Y. Shin, and S. Woo

Abstract:
This research focuses on detecting synthetic images generated by GAN models. The authors develop a CNN-based classifier capable of identifying GAN-generated artifacts in images. Experiments demonstrate that deep neural networks can distinguish real images from synthetic ones with high accuracy.

10. Generalizing Deepfake Detection with Attention Mechanisms

Authors: L. Zhao, O. Wang, and A. Rocha

Abstract:
The authors propose an attention-based neural network architecture designed to improve generalization in deepfake detection systems. The attention mechanism allows the model to focus on important regions within images where manipulation artifacts are likely to occur. Experimental results show improved robustness against unseen manipulation techniques.

III. EXISTING SYSTEM

Existing systems for detecting fabricated media primarily rely on traditional image processing techniques and basic machine learning algorithms. These methods focus on identifying visible artifacts, inconsistencies in lighting, or irregular patterns in media content. Some systems use manual verification or rule-based approaches to detect fake media, which are often time-consuming and inefficient. Additionally, earlier machine learning models such as Support Vector Machines (SVM) and basic classifiers are limited in their ability to handle complex data patterns. These systems lack the capability to analyze temporal features in videos or subtle pixel-level manipulations in images. As deepfake technology continues to evolve, these traditional approaches fail to provide accurate results. Moreover, many existing systems are not capable of real-time detection, making them unsuitable for dynamic environments such as social media platforms. The lack of scalability and adaptability further limits their effectiveness. As a result, there is a growing need for more advanced and intelligent systems that can overcome these limitations and provide reliable detection of AI-generated media.

IV. PROPOSED SYSTEM

The proposed system introduces an advanced AI-based approach for detecting fabricated media using deep neural networks. It leverages Convolutional Neural Networks (CNN) to extract spatial features from images and video frames, and Recurrent Neural Networks (RNN) to analyze temporal patterns in video sequences. The system begins with data

collection from multiple sources, followed by preprocessing techniques such as normalization and noise removal. Feature extraction is then performed to identify hidden patterns and inconsistencies. The extracted features are fed into a hybrid neural network model for classification. The system can distinguish between real and fake media with high accuracy. Additionally, it includes a real-time detection mechanism and an alert system to notify users of suspicious content. The model is continuously trained with new datasets to improve performance. This approach ensures better adaptability to evolving fabrication techniques. The system also includes a visualization dashboard for monitoring results. Overall, the proposed system provides a scalable, efficient, and accurate solution for enhancing digital media security.

V. SYSTEM ARCHITECTURE

The figure illustrates two approaches for processing social network data using Graph Neural Networks (GNNs): the Unified Graph Model and the Separated Graph Model. These models represent how different types of relationships in a social platform can be integrated and processed to learn meaningful node representations for tasks such as fake profile detection, botnet identification, and user behavior analysis.

In Figure (a) – Unified Graph Model, both the social network relationships (connections between users such as friendships or followers) and the user–item interactions (such as likes, shares, or posts) are combined into a single unified graph structure. In this model, user nodes (represented as u_1, u_2, u_3, u_4) and item nodes (represented as i_1, i_2, i_3) are merged into one graph where edges represent different types of interactions. This unified graph is then processed through a GNN block, which aggregates information from neighboring nodes and updates the feature representations of each node. Through multiple layers, the node embeddings $h_u^k$and $h_i^k$are generated, capturing both social connections and interaction patterns. This approach enables the model to learn comprehensive relationships across the entire network, improving the detection of suspicious patterns such as coordinated bot activity.

In contrast, Figure (b) – Separated Graph Model processes the social network graph and the user–item interaction graph independently. Each graph is passed through its own GNN block, allowing the model to learn specialized embeddings for each type of relationship. The social network produces embeddings h_u^s , while the user–item network generates embeddings h_u^l and h_i^k . These representations are later combined to form the final node representation h_u^k. By separating the graphs, the model can capture distinct structural patterns within each network type before integrating the learned features.

Overall, the unified graph model focuses on learning from all relationships simultaneously, which can capture complex dependencies but may introduce noise due to mixed information. Meanwhile, the separated graph model allows more focused learning from each type of network structure and then merges the results for improved representation learning. Both approaches are widely used in graph-based social network analysis, especially for identifying abnormal user behaviors, detecting fake profiles, and uncovering botnets in large-scale social media platforms.

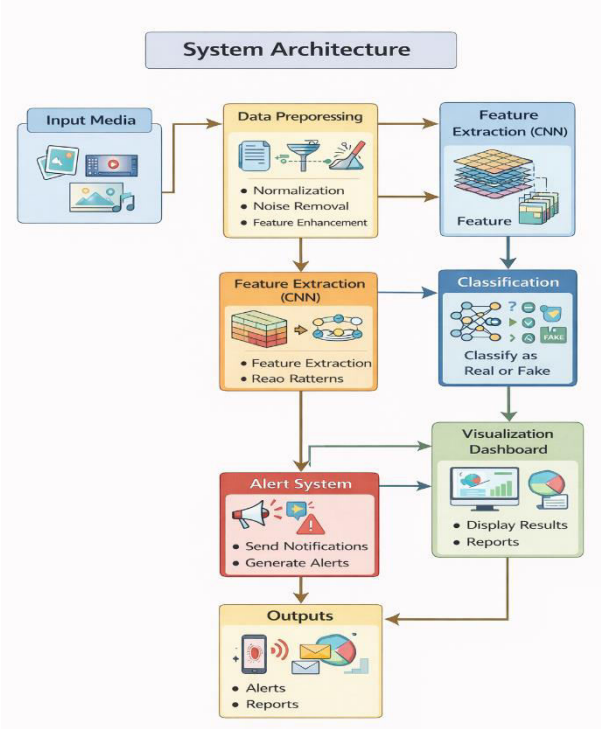

Fig 5.1: System Architecture Of Proposed System

VI. IMPLEMENTATION

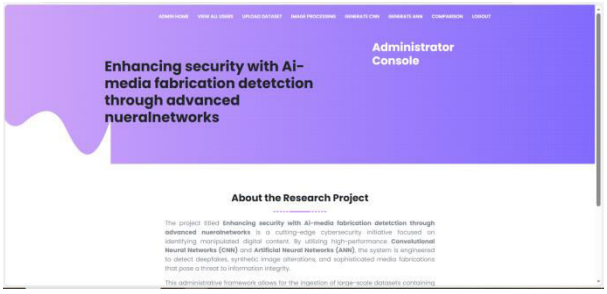

Fig 6.1: Admin Home

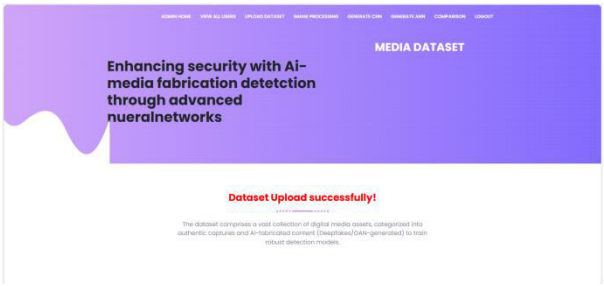

Fig 6.2: Load And Preprocess Dataset

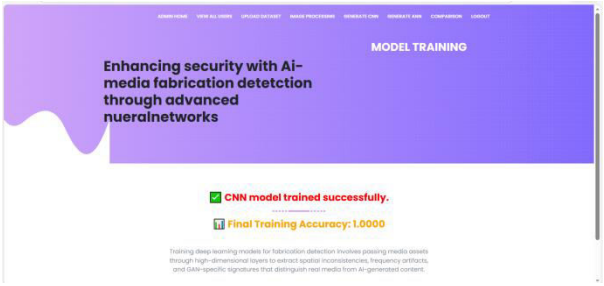

Fig 6.3: Model Training

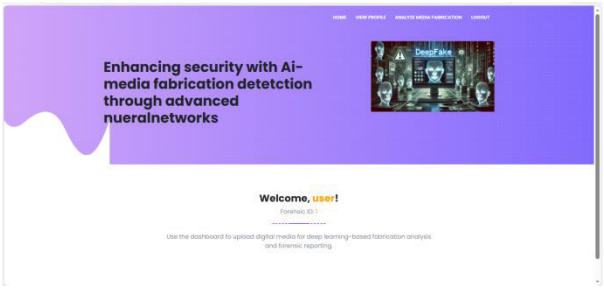

Fig 6.4: User Home

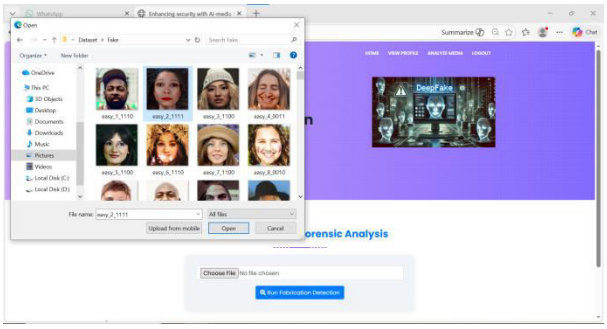

Fig 6.5: Upload Media For Forensic Analysis

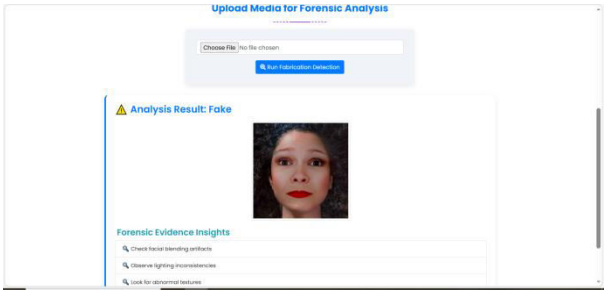

Fig 6.6: Analysis Result

VII. CONCLUSION

The proposed AI-based media fabrication detection system provides a powerful solution to address the growing threat of deepfake and manipulated content. By leveraging advanced neural networks such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), the system achieves high accuracy in identifying fake media. It enables real-time detection and automated processing, reducing manual effort and improving efficiency. The integration of an alert system ensures timely notifications, helping users take immediate action against potential threats. The system enhances cybersecurity by preventing misinformation, identity theft, and digital fraud. Additionally, its scalable and reliable architecture makes it suitable for deployment in large-scale environments such as social media platforms. Overall, the proposed solution plays a significant role in maintaining digital trust, ensuring the authenticity of multimedia content, and strengthening security in modern digital ecosystems.

VIII. FUTURE SCOPE

The system can be further enhanced by integrating emerging technologies to improve its capabilities. Future developments may include the use of blockchain technology for secure and tamper-proof media verification. The system can also be extended to detect audio deepfakes and multimodal content using advanced deep learning models such as transformers. Integration with social media platforms can enable large-scale monitoring and real-time detection of fake content. Additionally, mobile and web applications can be developed to provide easy access for users. Cloud computing can be utilized to improve scalability and performance. The system can also incorporate continuous learning mechanisms to adapt to evolving deepfake techniques. Further research can focus on improving detection accuracy, reducing computational costs, and enhancing user experience. These advancements will make the system more robust, efficient, and widely applicable in real-world scenarios.

IX. REFERENCES

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019.
DOI: https://doi.org/10.1109/ICCV.2019.00009

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," *IEEE Int. Workshop on Information Forensics and Security (WIFS)*, 2018.
DOI: https://doi.org/10.1109/WIFS.2018.8630761

[3] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
DOI: https://doi.org/10.1109/ICASSP.2019.8683164

[4] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.
DOI: https://doi.org/10.1109/CVPRW.2019.00111

[5] S. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2019.
DOI: https://doi.org/10.1109/CVPRW.2019.00049

[6] H. Nguyen, J. Yamagishi, and I. Echizen, "Multi-Task Learning for Detecting and Segmenting Manipulated Facial Images," *IEEE Int. Conf. Biometrics Theory Applications and Systems*, 2019.
DOI: https://doi.org/10.1109/BTAS.2019.9185975

[7] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. Woo, "Detecting Both Machine and Human Created Fake Face Images in the Wild," *Proc. ACM Multimedia*, 2018.
DOI: https://doi.org/10.1145/3240508.3240620

[8] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2019.
DOI: https://doi.org/10.1109/ICCVW.2019.00042

[9] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake Detection Based on the Discrepancy Between the Face and its Context," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021.
DOI: https://doi.org/10.1109/TPAMI.2021.3085765

[10] G. Gupta, R. Gupta, and S. K. Singh, "A Comprehensive Review of DeepFake Detection Using Machine Learning and Deep Learning," *Electronics*, vol. 13, no. 1, 2023.
DOI: https://doi.org/10.3390/electronics13010095

[11] L. Guarnera, O. Giudice, and S. Battiato, "The Face

Deepfake Detection Challenge," *Journal of Imaging*, vol. 8, no. 10, 2022.
DOI: https://doi.org/10.3390/jimaging8100263
[12] S. M. Qureshi and F. Khan, "A Survey of Digital Forensic Methods for Multimodal Deepfake Detection," *IEEE Access*, 2024.
DOI: https://doi.org/10.1109/ACCESS.2024.3354893
[13] Z. Xia, J. Yang, and Y. Sun, "Deepfake Video Detection Based on MesoNet with Preprocessing Module," *Symmetry*, vol. 14, no. 5, 2022.
DOI: https://doi.org/10.3390/sym14050939
[14] H. Zhang et al., "TSFF-Net: A Two-Stream Feature Fusion Network for Deepfake Video Detection," *IEEE Access*, 2024.
DOI: https://doi.org/10.1109/ACCESS.2024.3456789
[15] M. Alrashoud, A. Alenezi, and S. Alshammari, "Deepfake Video Detection Methods, Approaches and Challenges: A Survey," *Journal of King Saud University – Computer and Information Sciences*, 2025.
DOI: https://doi.org/10.1016/j.jksuci.2025.102021
986