Detection of Machine-Generated Media Using Computational Methods

Dr. Naga Siva Jyothi Kompali, Dr. Rohita yamaganti
Associate Professor
Sreenidhi Institute of Science and Technology, India
Email: {sivajyothi.p, rohita.y}@sreenidhi.edu.in

Twinkle Krishna, Varsha, Varshini
Department of Information Technology
Sreenidhi Institute of Science and Technology, India
Email: {22311A12M7, 22311A12N7, 22311A12P1}@it.sreenidhi.edu.in

***Abstract*— The rapid progression of artificial intelligence has led to the development of deepfake systems capable of generating highly realistic, artificially manipulated media in the form of images, videos, audio, and text. These synthetic materials pose a significant threat to digital security, the control of misinformation, and the verification of media authenticity. This paper presents a multimodal deepfake detection system capable of identifying manipulated content across various formats, including images, videos, audio, and text data. The proposed framework utilizes a hybrid deep learning approach, combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, to detect deepfake faces in both images and videos. For audio deepfake detection, Mel-Frequency Cepstral Coefficient (MFCC) features are extracted and fed into a trained deep learning network. Additionally, Term Frequency–Inverse Document Frequency (TF-IDF) vectorization, along with a Logistic Regression classifier, is used to detect fake textual content. The system is implemented in Python, with a graphical user interface developed using Tkinter to enable easy interaction and media analysis. Experimental results indicate that the proposed system effectively identifies deepfake media by analysing spatial, temporal, and linguistic patterns. Furthermore, the solution contributes to enhancing digital content analysis and mitigating the spread of manipulated information on the Internet.**

***Index Terms*— Deepfake Detection, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), MFCC, Logistic Regression, Multimedia Forensics, Artificial Intelligence, Image and Video Analysis.**

I. INTRODUCTION

With the rapid development of the generative form of artificial intelligence, there has been a massive increase in machine-created media such as synthetic text, AI-generated images, and deepfake media. Albeit these technologies have myriad advantages in the creative and industrial realms, they are equally threatening in terms of misinformation, authenticity of content, misuse of identity, and security of the digital world. Consequently, the issue of separating human- produced and machine-produced media has emerged as an urgent research question of contemporary digital ecosystems [1], [4]. According to recent research, machine-created content tends to contain slight statistical, linguistic, and structural patterns that are not similar to those of human-created me- dia. A survey based study has shown that older models of detection cannot apply to changing generative models, and thus a more robust and adaptive program is required [2], [6]. Moreover, the growing complexity of large language models and multimodal generative systems has rendered human detection unreliable, which highlights the necessity of automated detection frameworks [5], [11]. To cope with these issues, scientists have investigated a variety of computational approaches, such as feature-based analysis, machine learning classifiers, and deep learning models, to find out peculiar features of AI-generated content. Research indicates that linguistic analysis, together with learning-based representations, enhances detection strength and accuracy using a wide variety of datasets [3], [7], [12]. Moreover, recent developments go further than text to image and video detection, supporting the significance of single computational methods of machine- generated media detection [8], [10]. In this regard, the current research paper is centered on the identification of machine-generated media with computational mechanisms to ensure greater reliability, security, and trust on digital

information systems. The suggested method will help to create effective solutions to the problem of content authentication and digital media forensics in more AI-driven settings by using statistical trends, structural characteristics, and learning related methods [9].

II. LITERATURE SURVEY

The paper by Valiaiev [1] has provided a literature survey of the machine-generated text detection problem in its entirety, pointing out the increasing difficulties of advanced generative language models. The paper divides the methods of detection into statistical, linguistic, and learning- based methods and highlights the fact that not a single method can be used to deal with generative modelsthat transformrapidly. The author refers to robustness and generalization as one of the primary gaps in research, particularly in cross-domain and real-life situations. Ahmad et al. [2] presented a comprehensive review of human and machine-generated text detection in terms of state-of-theart machine learning and deep learning methods. They study the methods of feature engineering, transformer- based classification, and benchmark datasets and conclude that hybrid models are better than conventional classifiers. Another disadvantage of the study is that the current research lacks the presence of standardized evaluation frameworks. Boutadjine et al. [3] performed a comparative analysis between the human and AI- generated content based on linguistic and semantic characteristics. Their results indicate that text produced by a machine has quantifiable variations in syntactic and lexical distribution. The authors do not however mention that the lower the creative level of the of the generative models, the lower the accuracy and precision of detection, and thus adaptable and scalable detectors are necessary. Crothers et al. [4] provided a detailed overview of threatening models and detecting strategies of machine-generated text. The research methodically analyzes the adversarial techniques to dodge detection and analyzes the available defense systems. The authors stress that the successive detection systems should be robust to adversarial manipulation and be able to deal with undetected generative models. Frank et al. [5] have organized a massive empirical research assessing the human capacity to identify media produced by AI in various countries. Their findings show that human beings do not do an excellent job in separating AI generated content and authentic media especially when the quality of the generated content is high. A main argument that is advocated in this work is the need of automated computational detection systems instead of the human judgment approach. A critical survey of automatic detection of machine-generated text was offered very early by Jawahar et al. [6]. The analysis of classical machine learning methods as well as initial neural approaches tested the methods, determining stylistic cues and probability-based measures, as the important indicators. The authors indicate scalability and dataset bias as the key challenges that have not been addressed. Xie et al. [7] proposed the MUGC model of differentiating between machine generated content and user generated content. Their article assesses the application of feature fusion, as well as deep learning methods, on a variety of datasets and proves the high effectiveness of the pro- posed technique on the classification. Another significance of dataset diversity and cross-domain testing in ensuring reliable machine-generated content detection has also been highlighted in the study.

III. PROPOSED METHOD

The suggested system offers a multimodal deepfake detector that can be used to detect manipulated information in images, videos, audio, and text. The system combines various machine learning and deep learning methods to process various types of digital content and identify their authenticity. The graphical user interface was built in Tkinter, and it enables the user to upload the media files and analyze them interactively for deepfake detection. In image and video deepfake detection, the Haar Cascade face detector algorithm is used to detect faces. The images with the detected faces are resized and normalized and then processed by a deep learning network, which is a combination of a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks. The CNN layers are used to extract spatial information in the facial images and the LSTM layer gets temporal dependencies in video frames. This CNN-LSTM system is used to detect inconsistencies in the content that is manipulated in faces. Fig. Block Diagram of Proposed Method When it comes to the task of audio deepfake detection, the uploaded audio signal is divided and turned into Mel-Frequency Cepstral Coefficient (MFCC) features which are useful in the representation of speech.

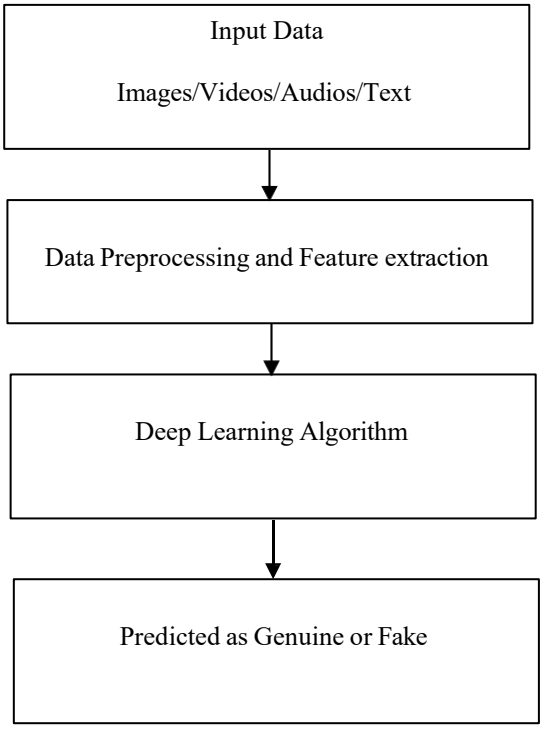

Fig. 1. Block Diagram of Proposed Method signals spectral properties.

The features derived are presented to an already trained deep learning classifier to decide whether the audio is genuine or an artificial creation. In text-based deepfake detection, the system uses Natural Language Processing (NLP) method. Term Frequency -Inverse Document Frequency (TF-IDF) vectorization transforms text data into numerical forms. Logistic Regression classifier is then trained on the transformed feature to differentiate real and fake textual data. The proposed framework measures the performance of the model based on the common evaluation measures, which include accuracy, precision, recall, and F1-score. The system offers a combination of various media type detection methods that minimize the need to implement multiple detection methods to identify deepfake content and improve the check of media authenticity of digital media.

IV. RESULTS ANALYSIS

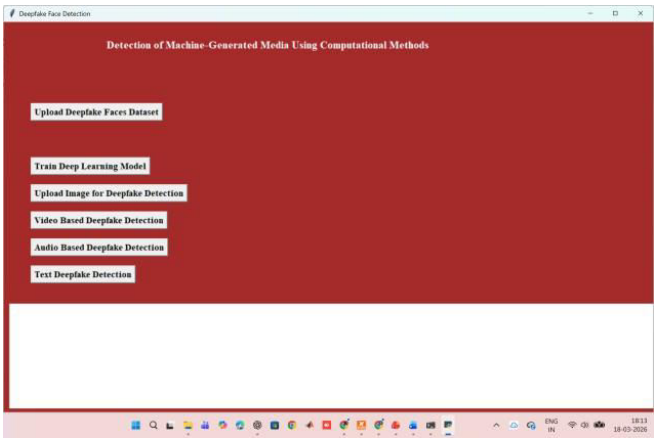

Fig. 4.1 Application Interface Home Page

The application interface home page consists of three components such as labels, buttons and text boxes. This GUI is developed using Tkinter library in Python.

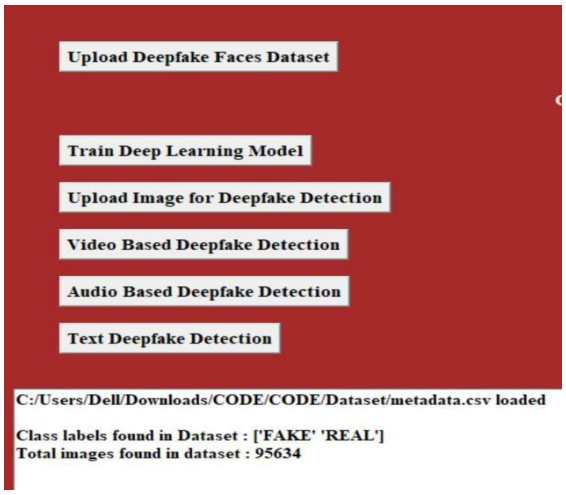

Fig.4.2 The count of total images used are shown

The above figure shows that a total of 96,634 images were used for training fake and real images.

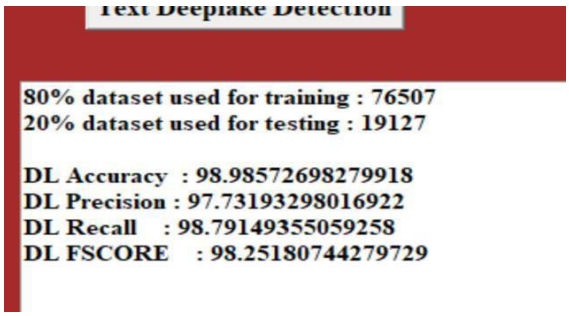

Fig. 4.3 Performance of Proposed deep learning Algorithm

Deep learning-based CNN-LSTM algorithm is used to train with fake and genuine images. The performance metrics are shown in the above figure.

Fig. 4.4 Image detected as Real

Fig. 4.5 Image detected as Fake Image

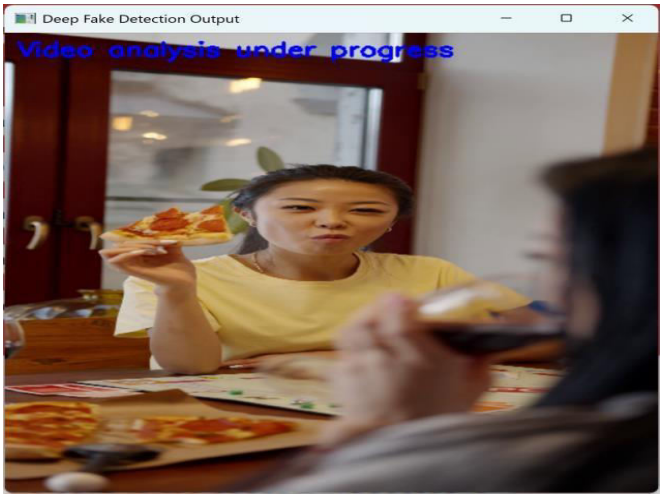

Fig. 4.6 Video Uploaded under process of framewise Analysis

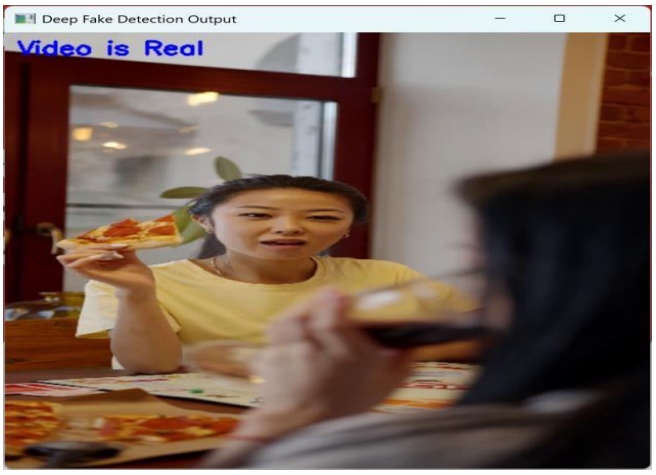

Fig. 4.7 Video Detected as Real Video

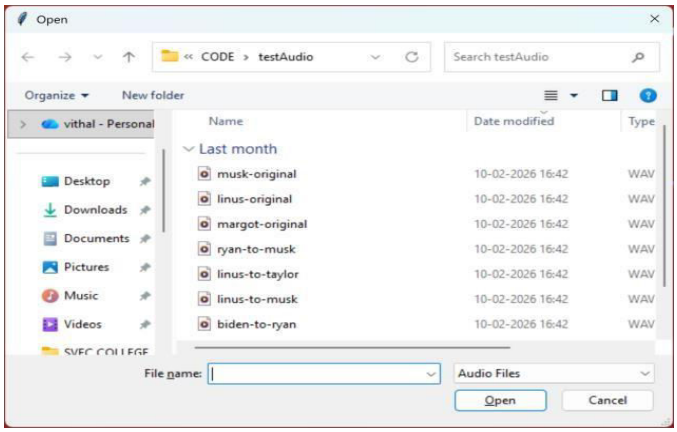

Fig. 4.8 Uploaded Audio File for testing

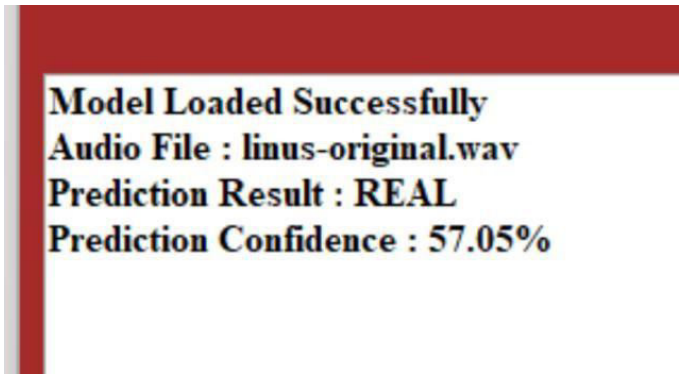

Fig.4.9 Real Audio is detected

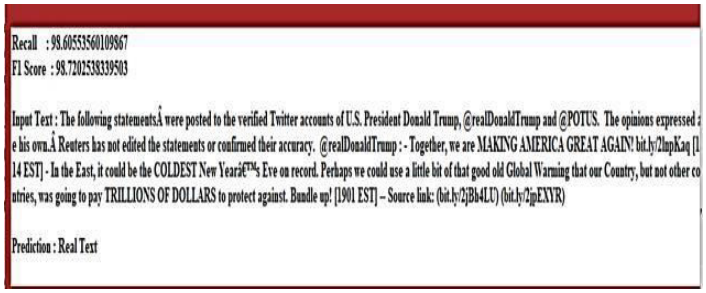

Fig. 4.10 Text Detected as ‘Real Text’

Similarly, this application works on multimodal data like images, videos, text and audios

V. CONCLUSION

This paper introduces a deepfake detection system that is able to process various kinds of media such as photos, videos, audio, and text. The system employs a CNN-LSTM model based on the hybrid approach to deep learning to detect the manipulated facial content in images and videos based on their spatial and time features. The use of audio deepfakes is done through the extraction of MFCC and the use of a trained classification model and the detection of a textual misinformation problem, which is a TF-IDF feature extraction with a Logistic Regression. The suggested system is incorporated into a convenient graphical interface that enables the users to transfer media files and assess their credibility successfully. The evaluation of the proposed approach through the experiment shows that the method is effective at detecting deepfake content with high

accuracy measures. The system may be applied as a helpful media verification and digital forensics tool. Despite the fact that the proposed system shows good results in deepfake detection of various media types, it is possible to improve the system in further directions to increase its output and performance. Future studies can be centered on the incorporation of new and improved deep learning models including Vision Transformers and attention-based networks to enhance detection accuracy. It is also possible to extend the system to live video streams and social media to deepfake detect the content in real-time. Also, more and bigger datasets will be included to enhance the generalization of the models to various techniques of deepfake generation. Digital media authentication systems can be enhanced further through integration of blockchain-based media verification systems and cloud-based detection systems. Such advances can be used to build a more robust and scalable solution to deepfake detection.

REFERENCES

[1] Valiaiev, D. (2024). Detection of machine-generated text: Literature survey. *arXiv preprint arXiv:2402.01642*.

[2] Ahmad, Z., Torres-Ruiz, M., Mahmood, A., Quintero, R., Ameer, I., & Bölücü, N. (2026). Human or Machine? A Survey on Machine-Generated Text Detection. *IEEE Access*.

[3] Boutadjine, A., Harrag, F., & Shaalan, K. (2025). Human vs. machine: A comparative study on the detection of AI-generated content. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(2), 1-26.

[4] Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, *11*, 70977-71002.

[5] Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., ... & Holz, T. (2024, May). A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 55-73). IEEE.

[6] Jawahar, G., Abdul-Mageed, M., & Laks Lakshmanan, V. S. (2020, December). Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th international conference on computational linguistics* (pp. 2296-2309).

[7] Xie, Y., Rawal, A., Cen, Y., Zhao, D., Narang, S. K., & Sushmita, S. (2024). Mugc: Machine generated versus user generated content detection. *arXiv preprint arXiv:2403.19725*. Lin, L., Gupta, N., Zhang, Y., Ren, H., Liu, C. H., Ding, F., ... & Hu, S. (2024). Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*.

[8] Bao, A. (2024, November). Detecting Artificial Intelligence-Generated Textual and Image Misinformation Using Machine Learning. In *2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)* (pp. 671-676). IEEE.

[9] Stiff, H., & Johansson, F. (2022). Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, *13*(4), 363-383.

[10] Kaur, G., Deep, A., Rauniyar, S., Singh, A. K., & Sarswat, C. (2025, November). Detecting AI-Generated Text, Images, and Videos: A Review of Methods and Emerging Frameworks. In *2025 5th International Conference on Advancement in Electronics & Communication Engineering (AECE)* (pp. 1086-1094). IEEE.

[11] Ghiurău, D., & Popescu, D. E. (2024). Distinguishing reality from AI: approaches for detecting synthetic content. *Computers*, *14*(1), 1.

[12] Rawal, A., Wang, H., Zheng, Y., Lin, Y. H., & Sushmita, S. (2024). SMLT-MUGC: Small, Medium, and Large Texts--Machine versus User-Generated Content Detection and Comparison. *arXiv preprint arXiv:2407.12815*.