IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Deepfake Image, Video And Audio Detection

Mahima A H
Dept. of Computer Science and Engineering
BNM Institute of Technology Bengaluru, India

Neha S
Dept. of Computer Science andEngineering
*BNM Institute of Technology*Bengaluru, India

Monica M
Dept. of Computer Science andEngineering
*BNM Institute of Technology*Bengaluru, India

Dr. Deepti Balaji Raykar
Associate Professor,
Dept. of Computer Science andEngineering
*BNM Institute of Technology*Bengaluru, India

Abstract**—** In the realm of digital media, the proliferation of deepfake technology has brought forth significant concerns regarding its potential for misuse, particularly in areas such as political disinformation and privacy infringement. To counter these threats, we propose a groundbreaking approach to deepfake analysis that encompasses both image and audio detection. This method integrates advanced techniques such as Res-Next CNN for precise frame-level feature extraction from images and MFCC feature extraction for audio, coupled with LSTM-based RNN for comprehensive temporal analysis across both modalities. This model exhibits a remarkable ability to accurately discern between authentic content and deepfakes in both images and audio recordings.

Additionally, our methodology emphasizes adaptability and scalability, ensuring its effectiveness across various digital platforms and evolving deepfake techniques in both visual and auditory domains. By continuously refining our model with ongoing advancements in AI and deep learning, we remain steadfast in our commitment to staying ahead of emerging threats posed by malicious actors.

In addition to its utility in detecting deepfakes across multiple modalities, our system features a user-friendly interface for swift mitigation of AI-generated manipulations. By leveraging the power of AI to combat AI, our methodology prioritizes simplicity and reliability, validated through rigorous evaluations on diverse datasets encompassing both image and audio samples. Through these efforts, we safeguard against the harmful impact of deepfakes, thus preserving the integrity of digital media and protecting individuals' privacy and societal well-being.

Keywords—AI-generated manipulations, DeepFake technology, Res-Next CNN, LSTM-based RNN, MFCC feature extraction

I. INTRODUCTION

Deepfake videos, epitomizing the pinnacle of digital manipulation, have sparked widespread concern in today's era due to their potential to deceive and manipulate audiences on an unprecedented scale. The proliferation of deepfake technology has engendered a myriad of challenges, notably in exacerbating misinformation and undermining trust in media content. These digitally manipulated videos, often indistinguishable from reality, pose significant risks by fuelling political manipulation, privacy breaches, and the spread of false narratives. However, technology plays a pivotal role in addressing these challenges. Advanced machine learning algorithms, such as deep neural networks, are harnessed to develop robust detection methods capable of distinguishing between genuine and deepfake images audios and videos with increasing accuracy. Additionally, automated content verification tools leverage artificial intelligence to identify inconsistencies and anomalies within media content, aiding in the rapid identification and removal of deepfakes from online platforms.

This research paper presents an innovative deep neural network-based framework and algorithm specifically crafted to discern deepfake videos within the realm of social media. Its primary contribution lies in addressing the escalating difficulty of differentiating between authentic and manipulated videos, a challenge exacerbated by the advancing sophistication of deep learning techniques used in the creation of deepfakes. By introducing a novel algorithm integrating a deep neural model, the paper aims to achieve highly accurate detection of deepfake videos disseminated across social media platforms. The overarching objective is to counteract the rampant spread of fabricated videos, which pose substantial risks to societal integrity, including threats to national security, democratic processes, and individual identities.

This endeavour is motivated by the exponential rise in face forgery within multimedia content over the past two decades, driven by the emergence of transformative technologies like deep learning networks capable of generating visually seamless deepfake videos. The capability to manipulate images and videos using AI-powered tools presents a formidable challenge to the authenticity of media content, potentially fuelling misinformation, undermining trust in journalistic sources, and precipitating broader societal repercussions. Thus, the urgent necessity for effective deepfake detection methodologies has become paramount in safeguarding against the adverse consequences of counterfeit videos across myriad facets of society.

II. LITERATURE SURVEY

This section delves into prior research concerning deepfake technology, with a focus on the proliferation of deceptive media content, concerns surrounding misinformation, and the potential societal impacts. By scrutinizing existing literature, the objective is to identify gaps, evaluate the efficacy of previous mitigation strategies, and lay the groundwork for our project's distinctive contributions. This review underscores the urgency for innovative approaches to combat the multifaceted challenges posed by deepfake technology in various domains, including politics, social media, and cybersecurity.

The study [1] focuses on creation of deepfakes which relies on the utilization of deep learning encoders and decoders, extensively employed in the realm of machine vision. Encoders function by extracting all features present in an image, which are then used by decoders to generate the fabricated image. Traditionally, generating deepfakes necessitated a substantial number of images and videos for training deep learning models, posing a significant challenge. However, in the current era, acquiring large datasets from social media platforms has become notably easier, leading to the development of more sophisticated deepfake techniques. Many deepfake algorithms leverage TensorFlow, an open-source software library initially developed by Google for internal use in machine learning and deep neural network research. TensorFlow has gained widespread popularity for various machine learning applications since its public release, offering a convenient means to design neural networks with satisfactory performance. Its APIs are compatible with the Python programming language, enabling easy experimentation with different Convolutional Neural Network (CNN) architectures and designs without extensive code modifications.

The paper [2] introduces a novel algorithm called the Self-Supervised Decoupling Network (SSDN), which integrates similarity decoupling techniques such as Siamese networks or contrastive learning. Furthermore, it utilizes Generative Adversarial Networks (GANs) to eliminate compression-related features. Implementation is anticipated to involve Python along with deep learning frameworks like TensorFlow or PyTorch, supplemented by OpenCV for preprocessing and managing image data.

The investigation in [3] unveils NA-VGG, a pioneering approach to detecting DeepFake face images, leveraging an enhanced VGG Convolutional Neural Network alongside image noise and augmentation methodologies. Initially, the RGB image undergoes processing via a SRM filter layer to accentuate image noise characteristics, subsequently serving as input for the network. Further enhancing the dataset, the noise image undergoes horizontal/vertical flipping for augmentation purposes.Despite these promising results, challenges loom regarding the model's sensitivity to variations in image noise, limited generalization due to dataset constraints, and reliance on the SRM filter, potentially hindering performance in the face of novel DeepFake techniques.

In the realm of deepfake video detection, the Certainty-based Attention Network (CAN) emerges as a groundbreaking solution, prioritizing certainty-key frames to make predictions with heightened confidence levels. Comprising two integral components, the CAN incorporates a certainty-based attention map generation segment and a certainty-attentive feature generation module. However, the effectiveness of such detection methods is contingent upon diverse and representative training datasets, with limitations arising from biased or limited data potentially compromising real-world applicability. Moreover, the CAN's adaptability to evolving deepfake techniques may be challenged by rapid advancements in the field, necessitating comprehensive coverage of emerging methods in the training data for robust detection capabilities has discussed in [4].

The proposed deepfake detection method in [5] leverages Convolutional Neural Networks (CNNs) to extract frame features and utilizes a custom classifier for identification purposes. Development likely involved Python, TensorFlow, or PyTorch for neural network implementation, with Keras facilitating model construction on these frameworks. Additionally, image and video processing tasks were performed using OpenCV, while GPU acceleration was achieved through the NVIDIA CUDA Toolkit to expedite computations. For interactive code execution, Jupyter Notebooks or Google Colab provided suitable environments. The effectiveness of CNN-based models is inherently reliant on the diversity and representativeness of the training dataset, with potential compromises in real-world generalization if the dataset is limited or biased.

Literature [6] presents the Deepfake Detection Model with Mouth Features (DFT-MF), employing a deep learning approach to identify Deepfake videos by analyzing lip/mouth movement. Utilizing Convolutional Neural Networks (CNNs), videos are segmented into frames and converted to grayscale images for processing and classification. The DFT-MF model utilizes supervised deep learning with CNNs to categorize videos as fake or real based on a threshold number of identified fake frames, determined by variables such as words per sentence, speech rate, and frame rate. The DFT-MF model faces limitations such as it primarily focuses on the mouth region, potentially overlooking manipulations in other facial or contextual cues, and its efficacy heavily relies on accurate lip/mouth movement detection, making it sensitive to variations in facial expressions, lighting conditions, or video quality.

III. PROPOSED SOLUTION

The proposed solution for deepfake image and video detection involves a comprehensive approach, integrating data analysis, system design, and model development. Thorough analysis highlights the importance of balanced training datasets to mitigate bias and variance in algorithmic predictions [7]. Leveraging deep learning frameworks like PyTorch [8], the system architecture is designed with a focus on feature extraction and classification, utilizing a combination of MTCNN (Multi-Task Cascaded Convolutional Networks) for face detection [9], FaceSwap[14] for facial manipulation detection, and ResNeXt for feature extraction from images and videos. Long Short-Term Memory (LSTM) networks are incorporated to capture temporal dependencies in video sequences.

Preprocessing steps involve image segmentation, face detection, and frame extraction, ensuring uniformity and relevance of data inputs. By incorporating ensemble methods and fine-tuning techniques, the model aims to enhance detection accuracy and adaptability to evolving deepfake techniques. The user interface, developed using Django framework [10], offers seamless interaction, allowing users to upload images and videos and receive real-time predictions with confidence scores, thereby facilitating efficient deepfake detection.

The utilization of the librosa library[11] for audio processing within the Python environment. Librosa provides convenient tools for loading audio files and extracting relevant features necessary for our deepfake detection system. In the provided code snippet, librosa is employed to load the audio file and extract MFCC features, which serve as essential inputs for the classification model.

The aim of this proposed solution is to develop a robust deepfake detection system that outperforms existing methods by leveraging advanced techniques and addressing limitations such as biased training data and evolving manipulation techniques.

System Architecture:

The following Figure1 portrays the proposed system's architecture with which the system is built, and hence can be used to detect deepfake images and videos.

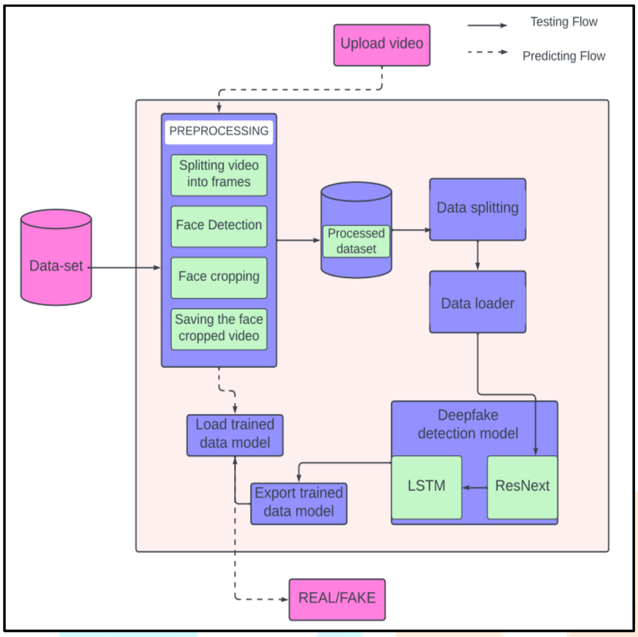

FIGURE 1. System Architecture

In this cloud-based system, users upload both real and fake datasets comprising images and videos to secure cloud storage services such as AWS S3, Google Cloud Storage, or Azure Storage Blobs. Upon upload, the datasets undergo preprocessing steps including noise reduction, image enhancement, and format standardization to ensure consistency and quality. The preprocessed datasets are then split into training, validation, and testing sets, with a data loader component retrieving batches of data during model training and evaluation[12]. For image data, the ResNeXt architecture is employed for feature extraction, capturing discriminative information for classification of real and fake content.

Meanwhile, Long Short-Term Memory (LSTM) networks are utilized for video classification, capturing temporal dependencies in video sequences[13]. Evaluation of both image and video classification models is performed using metrics such as accuracy, precision, recall, and F1-score, with results logged and stored for analysis. The system includes user authentication mechanisms like AWS Cognito or Azure Active Directory, ensuring secure access to data and functionalities. Visualization tools present evaluation results and performance metrics via dashboards, providing users with insights into model effectiveness and performance over time .

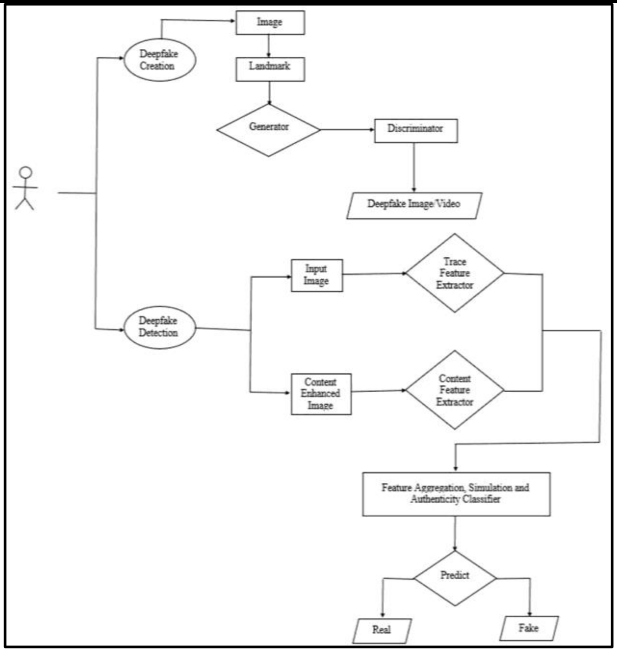

FIGURE 2. Use case diagram

Upon initiating the detection process, the system performs feature extraction and classification using advanced techniques like Res-Next CNN and LSTM-based RNN. Users receive real-time predictions accompanied by confidence scores, providing insights into the likelihood of the content being authentic or deepfake. Based on these results, users can take appropriate actions, such as content removal or further investigation.

IV.WORKFLOW

In our quest to tackle the deepfake menace, we begin by gathering diverse datasets and preprocessing them meticulously. Next, we design a robust model architecture, leveraging advanced techniques like CNNs and LSTMs. With a focus on optimization and validation, we fine-tune our model and evaluate its performance rigorously. Comparative analysis with existing methods guides our path towards innovation, while seamless deployment and integration ensure real-world impact. This workflow embodies our commitment to combating deepfake proliferation and upholding the integrity of digital content.

1. Data Collection and Preprocessing:

- Diverse datasets containing real and deepfake videos are collected from sources like Kaggle.
- Preprocessing involves standardizing formats, removing noise, and extracting features like facial landmarks using OpenCV.
- Videos are split into frames, and irrelevant information is discarded, focusing only on facial regions using techniques like face detection.
- Librosa[11] plays a crucial role in our deepfake detection system by providing the necessary tools for audio processing, enabling us to extract informative features essential for accurate classification.

2. Model Architecture Design:

- The architecture comprises a combination of Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) for temporal analysis.

- ResNext CNN models are utilized for feature extraction such as eyes, ears, nose, chin etc. from video frames, leveraging pre-trained models for efficient processing.
- Long Short-Term Memory (LSTM) networks are integrated to capture temporal dependencies between frames, enhancing the model's ability to discern subtle manipulations.

3. Model Training and Optimization:

- Training the model involves fine-tuning the CNN and LSTM layers using PyTorch framework and TensorFlow, with CUDA support for GPU acceleration.
- Hyperparameters such as learning rate, batch size, and dropout rates are optimized through iterative experimentation.
- Data augmentation techniques like flipping, rotation, and scaling(resizing of the images) are applied to augment the dataset and prevent overfitting.

4. Evaluation:

- The trained model's performance is evaluated using validation datasets, assessing metrics like probability, precision.
- Statistical tests validate the significance of the results and verify the model's ability to generalize to unseen data using probability.

5. Comparison with Existing Methods:

- The proposed model is benchmarked against baseline methods, including traditional machine learning algorithms and existing deepfake detection frameworks.
- Comparative analysis considers performance metrics, computational efficiency, and scalability, highlighting the superiority of the proposed approach.
- Qualitative assessments examine interpretability, ease of implementation, and potential for real-world deployment.

6. Deployment and Integration:

- The trained model is deployed in real-world scenarios through user-friendly interfaces developed using Django frameworks.
- Extensive testing ensure compatibility, reliability, and scalability of the deployed system, with provisions for continuous monitoring and updates.

V. ADVANTAGES

1. Combatting Misinformation:

Deepfake detection helps in identifying and mitigating the spread of misinformation and fake news, especially on social media platforms. By identifying manipulated content, users can avoid being misled by false information.

2. Preserving Authenticity:

Deepfake detection helps in preserving the authenticity of digital media by distinguishing between genuine and manipulated content. This is crucial for maintaining trust in online information and media.

3. Protecting Privacy:

Deepfake detection can help protect individuals' privacy by identifying and flagging unauthorized manipulations of their images or videos. This is particularly important in preventing the misuse of personal data for malicious purposes.

4. Preventing Fraud and Cybercrime:

Deepfake detection can assist in preventing various forms of fraud and cybercrime, such as identity theft, phishing attacks, and financial scams. By identifying fake videos or images used for fraudulent purposes, potential victims can be alerted and protected.

5. Maintaining Trust in Digital Content:

Deepfake detection contributes to maintaining trust in digital content, including media, entertainment, and advertising. By ensuring that content is authentic and unaltered, consumers can have confidence in the integrity of the information they consume.

6. Facilitating Content Moderation:

Deepfake detection aids content moderation efforts on online platforms by identifying and removing harmful or inappropriate content. This helps in creating safer and more responsible online environments for users.

Overall, deepfake image and video detection play a crucial role in safeguarding the integrity of digital media, protecting individuals' privacy, combating misinformation, and supporting various aspects of cybersecurity and law enforcement.

VI. RESULTS

- **Deepfake Image Creation:**

For the creation of deepfake image a source image and a destination image are used. Some of the features of source images are extracted and swapped with the features of destination image. The destination image will the deepfake image created.

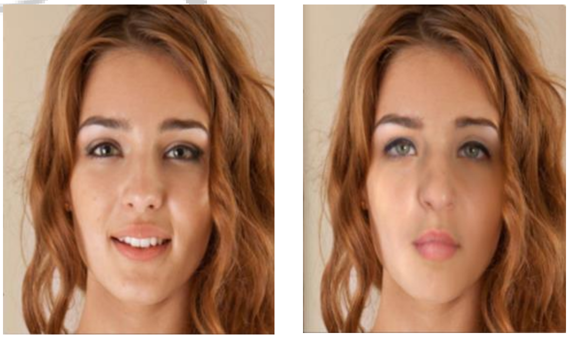

FIGURE 3. Deepfake Image Creation

- **Deepfake Image Detection:**

Upon processing an image through our deepfake detection system, the output includes a probability score indicating the likelihood of the image being fake or real. This probability score serves as a quantitative measure of confidence in the prediction, with values ranging from 0% to 100%. For instance, if the system assigns a probability of 50% to an image, it signifies a balanced uncertainty regarding its authenticity, suggesting an equal chance of it being either real or fake. A higher probability score, such as 80% or above, indicates stronger confidence in the classification, leaning towards either real or fake, depending on the direction of the probability.

Conversely, a lower score, such as below 50%, suggests a higher degree of uncertainty or ambiguity in the classification. This probabilistic output empowers users to make informed decisions based on the level of confidence in the authenticity of the image, enabling effective identification and mitigation of deepfake content in various contexts, including media forensics, cybersecurity, and content moderation.

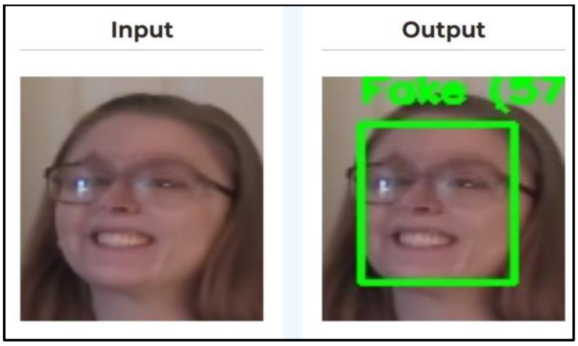

FIGURE 4. Fake Image Detection

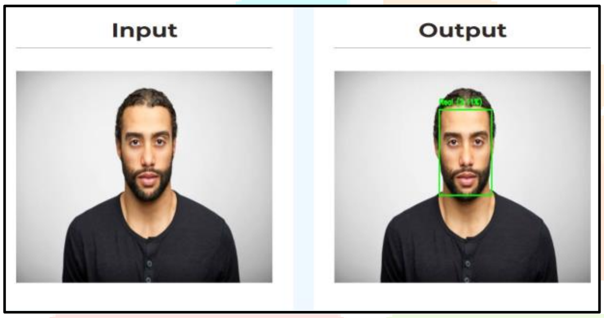

FIGURE 5. Real Image Detection

- **Deepfake Audio Detection:**

Upon processing each audio file, the system efficiently extracted relevant features using the librosa library and subsequently employed a Random Forest classifier trained on a diverse dataset encompassing both real and fake audio instances. The classification results demonstrated the system's capability to discern between real and fake audio content with notable accuracy.
Librosa's functionality allows to handle audio files efficiently, ensuring accurate feature extraction while optimizing computational resources. By leveraging librosa, we can streamline the preprocessing pipeline and focus on extracting discriminative features that facilitate the classification of real and fake audio content.
For instance, when testing the system on a sample audio file, it correctly classified the content as 'Real' or 'Fake' based on the extracted features and learned decision boundaries. This successful classification underscores the effectiveness of the employed methodology, leveraging feature extraction techniques and machine learning algorithms to distinguish between authentic and manipulated audio content, thereby contributing to the broader effort of combating the proliferation of deepfake media.

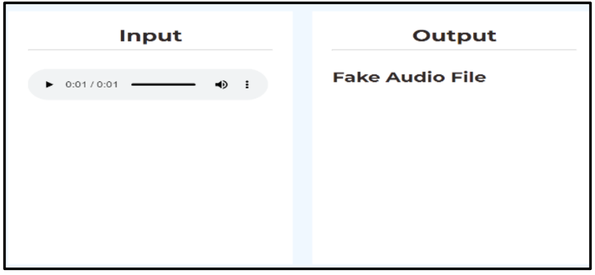

FIGURE 6. Fake Audio File Detection

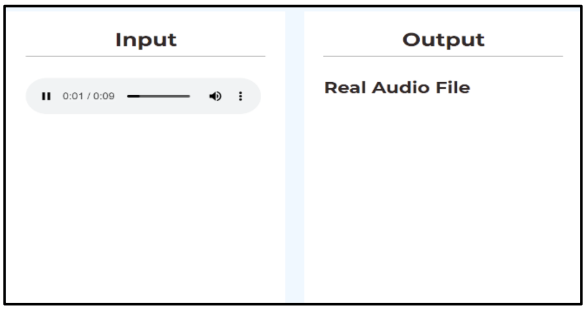

FIGURE 7. Real Audio File Detection

VII. CONCLUSION

The proposed approach to deepfake analysis, integrating advanced techniques such as Res-Next CNN and LSTM-based RNN for both image and audio detection, presents a robust solution to combat the proliferation of deepfake technology. By leveraging diverse datasets and continuously refining our model with advancements in AI and deep learning, we've developed a system capable of accurately discerning between authentic content and deepfakes across multiple modalities.
The methodology's adaptability and scalability ensure its effectiveness in detecting deepfakes on various digital platforms and against evolving manipulation techniques. Moreover, the user-friendly interface facilitates swift mitigation of AI-generated manipulations, contributing to the preservation of digital media integrity and protection of individuals' privacy and societal well-being.
Furthermore, this approach employs probability as a measure of confidence in the classification rather than solely relying on accuracy. This is advantageous because probability provides a nuanced understanding of the model's certainty in its predictions. While accuracy measures the overall correctness of predictions, probability allows users to gauge the level of confidence in individual classifications. This nuanced approach to evaluation enhances the reliability and interpretability of our deepfake detection system, empowering users to make informed decisions based on the confidence levels associated with each prediction.
In essence, this comprehensive methodology, coupled with the utilization of probability for confidence estimation, underscores our commitment to combatting the harmful effects of deepfake technology and safeguarding the integrity of digital media in the face of emerging threats.

VIII. FUTURE ENHANCEMENT

Several enhancements can elevate deepfake detection system. Integrating multimodal fusion techniques will improve accuracy by combining information from sources like images and audio. Adversarial training methods can fortify the model against sophisticated deepfake generation techniques. Dynamic adaptation to emerging deepfake techniques is vital for sustained effectiveness in real-world scenarios. Human-in-the-loop systems can enhance performance by incorporating human feedback into the detection process.

Privacy-preserving techniques should be explored to handle sensitive data during detection. Enhancing the explainability of the model's decisions fosters user trust and understanding. Optimizing the system for real-time detection is crucial for proactive mitigation of harmful content. By exploring these avenues, we can advance our deepfake detection system's capabilities, ensuring its effectiveness against evolving threats.

REFERENCES

[1] Hady A. Khalil, Shady A. Maged, "Deepfakes Creation and Detection Using Deep Learning", Department of Mechatronics Engineering, Ain Shams University, Cairo, Egypt, June 2021.

[2] Jian Zhang, Jiangqun Ni and Hao Xie, "Deepfake videos detection using self-supervised decoupling network", School of Computer Science and Engineering, Sun Yat-sen University, China, June 2021

[3] Xu Chang1,2, Jian Wu1,2, Tongfeng Yang1, Guorui Feng1, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network", School of Cyber Security, Shandong University of Political Science and Law, Jinan, July 2020.

[4] Dae Hwi Choi, Hong Joo Lee, Sangmin Lee, Jung Uk Kim, and Yong Man Ro, "Fake video detection with certainty-based attention network", Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea, October 2020.

[5] Alakananda Mitra, Saraju P. Mohanty, Elias Kougianos University of North Texas, USA, Peter Corcoran National University of Ireland, Galway, Ireland, "A Novel Machine Learning based Method for Deepfake Video Detection in Social Media" ,2020.

[6] Mousa Tayseer Jafar Mohammad Ababneh Mohammad Al-Zoube Ammar Elhassan, "Digital Forensics and Analysis of Deepfake Videos", Princess Sumaya University for Technology Amman, Jordan, 2020.

[7] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11).M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, ''Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,'' 2021, arXiv:2103.00484.

[8] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems (pp. 8024-8035).

[9] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499-1503.

[10] Holovaty, A., & Kaplan-Moss, J. (2009). The definitive guide to Django: Web development done right. Apress.

[11] https://www.javatpoint.com/librosa-library-in-python.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[13] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[14]