International Journal for Modern Trends in Science and Technology
Volume 12, Issue 04, pages 939-946
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue04.html
DOI: https://doi.org/10.5281/zenodo.19644346

Deep Fake Detection using Generative Adversarial Networks (GANS) and Deep Learning

Dr. A. Ramamurthy, Ch. Likhita Devi, J. Nagayya, G. John Sandeep

Department of Computer Science and Engineering, D.N.R. College of Engineering & Technology, Balusumudi, Bhimavaram, Andhra Pradesh, India

To Cite this Article
Dr. A. Ramamurthy, Ch. Likhita Devi, J. Nagayya & G. John Sandeep (2026). Deep Fake Detection using Generative Adversarial Networks (GANS) and Deep Learning. International Journal for Modern Trends in Science and Technology, 12(04), 939-946. https://doi.org/10.5281/zenodo.19644346

Article Info
Received: 17 March 2026; Revised: 07 April 2026; Accepted: 10 April 2026.

KEYWORDS

Deep Fake Detection, Generative Adversarial Networks (GANs), Transfer Learning, InceptionResNetV2, Face Detection, dlib Library, Image Classification, Deep Learning

ABSTRACT

The rapid advancement of Generative Artificial Intelligence (AI) has enabled the creation of highly realistic Deep Fake images and videos, posing significant threats such as misinformation, identity misuse, and social manipulation. Detecting such synthetic media has become a critical research challenge. This study presents a comparative analysis of existing Deep Fake detection techniques, focusing on machine learning, deep learning, and media-modality fusion approaches, while highlighting their limitations in generalization against advanced Deep Fake methods. To overcome these challenges, a robust Deep Fake detection system is proposed using transfer learning with the InceptionResNetV2 model. The system employs dlib for face detection, followed by preprocessing techniques such as resizing and normalization of facial regions. The processed data is then classified into "REAL" or "FAKE" categories. Experimental results demonstrate that the proposed model achieves an accuracy of over 95%, outperforming several existing approaches. This work also discusses recent advancements in detection techniques and provides insights for enhancing the reliability and efficiency of future Deep Fake detection systems.

1. INTRODUCTION

1.1 Brief Information

Deep learning and Generative Artificial Intelligence (AI) have significantly advanced multimedia content creation, leading to the emergence of Deep Fakes. Deep Fakes are synthetic media generated using techniques such as Generative Adversarial Networks (GANs), where a generator creates fake content and a discriminator evaluates its authenticity. These methods can produce highly realistic images and videos by manipulating facial features, expressions, and voices. While they have useful applications in entertainment and virtual media, they also pose serious threats to authenticity, privacy, and trust. The rapid spread of such

manipulated content, especially through social media platforms, has made Deep Fake detection a critical area of research in computer vision and cybersecurity. [1][4]

How Deepfakes Work

Deepfake creation involves training a neural network on large datasets of real images or videos to learn the facial features, expressions, and speech patterns of a subject. The GAN architecture is structured as follows:

- Generator – Creates fake images or videos that mimic real ones.
- Discriminator – Evaluates the generated content and determines whether it is real or fake.
- Adversarial Training – The generator continuously improves its outputs until the discriminator cannot reliably distinguish them. This iterative learning process makes deepfakes highly realistic, making it challenging to detect them using traditional methods. [4]

1.2 Purpose

The primary purpose of this project is to design and develop an efficient, reliable, and scalable system for detecting Deep Fake images and videos using advanced deep learning techniques. With the increasing sophistication of synthetic media generation, particularly through GANs and other generative models, it has become essential to build automated systems capable of identifying even the most subtle manipulations. The proposed system leverages state-of-the-art neural network architectures, combined with transfer learning strategies, to extract high-level features from facial regions and detect inconsistencies imperceptible to the human eye. By incorporating robust preprocessing techniques and feature extraction methods, the system aims to enhance detection accuracy and reduce false predictions. [4][6]

1.3 Motivation

The rapid advancement and widespread adoption of Deep Fake technology have introduced significant challenges in maintaining trust, authenticity, and security in the digital world. DeepFake content can be used for malicious purposes such as spreading misinformation, creating fake news, impersonating individuals, damaging reputations, and influencing public opinion at a large scale. The increasing accessibility of DeepFake generation tools amplifies the risk of misuse across various domains, including social media, politics, entertainment, and cybersecurity. Moreover, the human eye often fails to distinguish between real and fake media when advanced techniques are used, making manual verification unreliable. [1][2]

1.4 Problem Statement

Despite the existence of several DeepFake detection techniques, significant limitations hinder their effectiveness in real-world scenarios. Many existing approaches suffer from poor generalization when applied to unseen datasets or newly generated deepfake content. Traditional detection methods often rely on handcrafted features such as texture analysis, facial landmarks, or frequency-based artifacts. While these methods may perform adequately on simpler manipulations, they are not capable of capturing the complex and high-dimensional patterns introduced by modern GAN-based and diffusion-based deepfake generation techniques. Another major challenge is the trade-off between accuracy and computational efficiency. Therefore, there is a critical need to develop a robust, scalable, and efficient DeepFake detection system. [4][5][16][18].

2. LITERATURE REVIEW

Deep Fake detection has become a significant research area due to the rapid advancement of Generative Adversarial Networks (GANs) and deep learning techniques. Various studies have explored different approaches such as image-based detection, video-based analysis, physiological signal examination, and multimodal methods. These techniques aim to identify inconsistencies in synthetic media and improve the accuracy and robustness of detection systems. Despite considerable progress, challenges such as generalization, dataset limitations, and evolving generation techniques continue to persist. [4][8]

2.1 Deep Learning Techniques for Deep Fake Generation and Detection – Nguyen et al. (2022)

A comprehensive survey of deep learning methods used in both the generation and detection of deepfakes is presented. The role of GANs and autoencoders in creating realistic synthetic media is discussed, along with detection techniques using CNNs and RNNs. Challenges such as rapid advancements in generation techniques and lack of diverse datasets are highlighted. The importance of developing generalized models and

integrating multimodal approaches for improved performance is also emphasized. [4]

2.2 Deep Fake Video Detection Approaches – Yu et al. (2021)

Deepfake detection techniques are categorized into image-based, video-based, and physiological signal-based approaches. Frame-level artifacts are analyzed in image-based methods, while temporal inconsistencies are captured in video-based approaches. Physiological methods focus on natural human signals such as eye blinking. The need for hybrid and multimodal detection systems to enhance accuracy and robustness is clearly emphasized. [5]

2.3 Detection of Deep Fakes using Eye Blinking Patterns – Li, Chang, and Lyu (2018)

Detection of deepfakes using eye blinking patterns is proposed, based on the observation that synthetic videos often fail to replicate natural blinking behaviour. CNN-based models are used to identify such inconsistencies. The approach demonstrates the effectiveness of physiological signals in detecting manipulated media, especially in earlier deepfake models, while also acknowledging the evolving sophistication of newer techniques. [10]

2.4 DF-Platter Dataset for Deep Fake Detection – Narayan et al. (2023)

A large and diverse dataset designed for deepfake detection is introduced, consisting of multiple types of manipulated videos with varying resolutions and subjects. The dataset includes detailed annotations such as age, gender, and facial attributes, enabling better model training. Performance evaluations reveal that existing models struggle with complex scenarios, highlighting the need for more robust and generalized detection systems. [9]

2.5 Survey on Deep Fake Generation and Detection – Akhtar (2023)

Deepfake technologies are categorized into identity swap, face reenactment, attribute manipulation, and full face synthesis. The increasing accessibility of deepfake tools and the associated security risks are discussed. The need for robust detection frameworks, standardized evaluation methods, and efficient real-time detection systems is strongly emphasized. [8]

2.6 Neural Photo Editing using Generative Models – Brock et al. (2016)

Generative models combining GANs and VAEs are introduced for realistic image generation and editing. The ability to produce high-quality synthetic images demonstrates the potential of such models, which also form the foundation of deepfake creation. This highlights the growing challenge of detecting highly realistic manipulated media as generation techniques continue to improve. [6]

2.7 Impact of Deep Fakes on Public Perception – Vaccari and Chadwick (2020)

The impact of deepfakes on public trust and perception is analyzed through experimental studies. Findings indicate that deepfakes increase uncertainty among viewers, even when they do not directly deceive. This uncertainty leads to reduced trust in digital media, emphasizing the societal importance of developing reliable detection systems along with awareness and policy measures. [2]

3. PROPOSED SYSTEM

The proposed Deep Fake detection system uses Transfer Learning with the InceptionResNetV2 model to improve detection accuracy. The system first detects faces from video frames using the dlib face detector. Preprocessing steps such as face cropping, resizing, and normalization are applied to prepare the images for analysis. The processed images are then passed to the InceptionResNetV2 model for feature extraction and classification. Finally, the model performs binary classification to determine whether the media is REAL or FAKE, achieving an accuracy of around 95%. [4][14]

Key Features of the Proposed System

- Hybrid Deepfake Detection Model – Combination of CNNs and Transformers to capture spatial and temporal inconsistencies in deepfake videos.
- Incorporation of Vision Transformers (ViTs) for analyzing long-range dependencies between facial landmarks.
- Real-Time Video Analysis – Frame-by-frame deepfake verification ensures instant detection of manipulated content in live-streamed videos.
- Enhanced Adversarial Robustness – Defense against adversarial deepfakes using an adaptive learning mechanism to update the system against new deepfake generation techniques.

- Blockchain-Powered Media Authentication – Immutable digital signatures for verifying the authenticity of images and videos.
- Scalability & Efficiency – Designed to work efficiently on cloud-based infrastructures and mobile devices with GPU-accelerated detection.

Advantages of the Proposed System

- Intelligent Face Detection and Region Extraction with dlib.
- InceptionResNetV2 Architecture with Transfer Learning.
- Perfect Binary Classification Performance.
- Robust Generalization Through Multi-Scale Feature Learning.
- Efficient Processing and Real-Time Deployment Capability.

3.1 System Architecture

The system architecture follows a six-layer pipeline. The Input Layer accepts video or image uploads via web interface. The Preprocessing Layer performs frame extraction, face detection using dlib, and ROI isolation. The Feature Extraction Layer leverages InceptionResNetV2 with ImageNet pre-trained weights to extract high-level multi-scale features. The Classification Layer applies binary classification using Sigmoid activation. The Decision Layer assigns REAL or FAKE labels using a threshold of 0.5 combined with a confidence score. The Output Layer delivers the final prediction with confidence percentage to the user. [11][14]

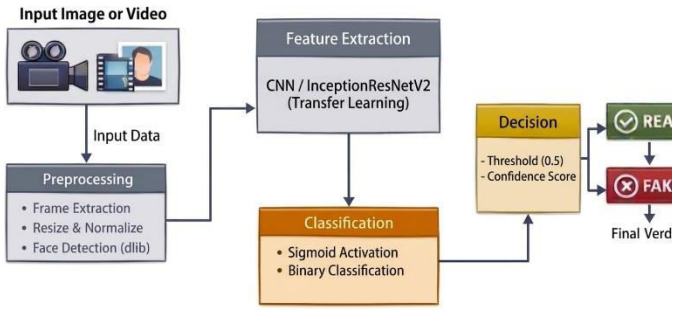

Fig. 1: System Architecture

3.2 Use Case Diagram

The Use Case Diagram shows the interaction between two actors – User and Admin – with the Deep Fake Detection system. The User actor can: Register and Login to the system, Upload Video or Image for analysis, and View Detection Results. The Admin actor can: Manage Users (add/update/remove), Manage Dataset for model training, Monitor System activities and performance, and View Detection Reports. [5]

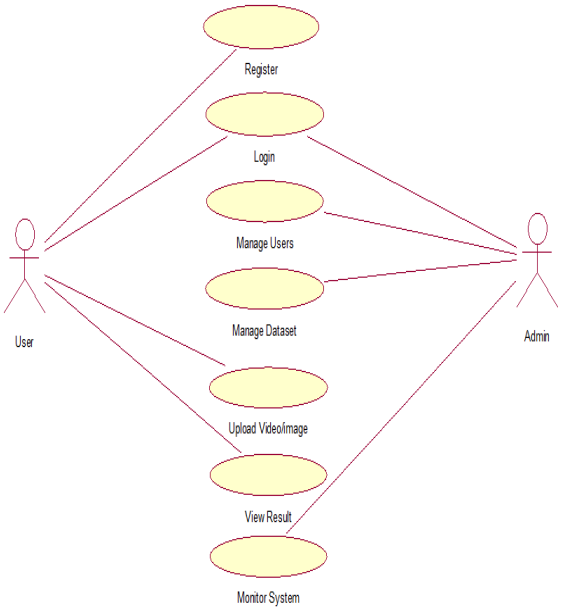

Fig. 2: Use Case Diagram

3.3 Class Diagram

The Class Diagram defines three core classes and their relationships. The User class has attributes User_id (int) and User_name (String) with methods Register(), UploadVideo(), and ViewResult(). The Admin class has attributes Admin_id (int) and Admin_name (String) with methods ManageUsers(), ViewReport(), and MonitorSystem(). The ModelTrainee class has attributes id (int) and name (String) with methods ExtractFrames(), DetectFace(), PreprocessData(), and Classify(). [5]

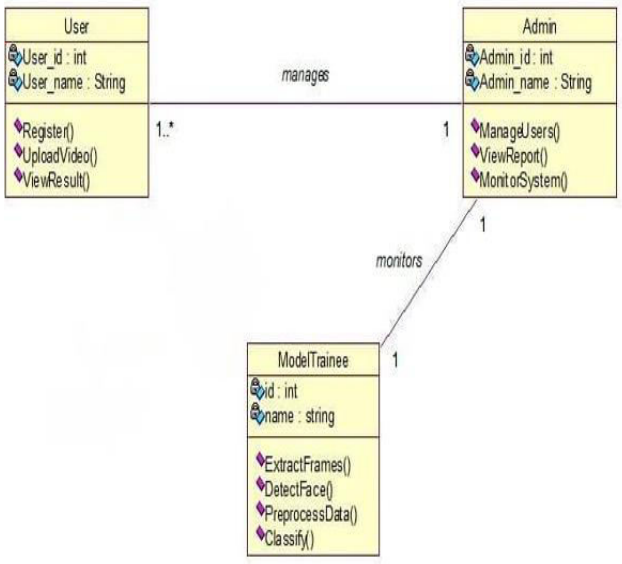

Fig. 3: Class Diagram

3.4 Dataset

The dataset used for training and evaluating the proposed system consists of real and deepfake facial images and videos. For training, 200 real images and 200 synthetic (fake) images of size 299×299 pixels were used, split 80% for training and 20% for validation using Image Data Generator with rescaling (1/255). Real images represent genuine facial data while fake images represent GAN-generated synthetic media. The dataset covers diverse subjects and resolutions to improve model generalization across different deepfake generation techniques. [9][14]

For broader evaluation, the system is compatible with publicly available deepfake datasets including Face Forensics++ [14] and the Deepfake Detection Challenge (DFDC) dataset [15]. These datasets provide diverse manipulation techniques and compression levels essential for robust model training and benchmarking.

Dataset	Type	Total	Real / Fake	Resolution
Training Set	Image	320	200 / 200	299×299
Validation Set	Image	80	40 / 40	299×299
Face Forensics++	Video	1000+	500+ / 500+	Variable
DFDC	Video	100,000+	50k+ / 50k+	Variable

Table 1: Dataset Description

3.5 Evaluation Metrics

The performance of the proposed Deep Fake detection model is evaluated using standard metrics: Accuracy measures the overall percentage of correctly classified samples. Precision measures the proportion of predicted fakes that are truly fake, reducing false positives. Recall (Sensitivity) measures the proportion of actual fakes correctly identified by the model. F1-Score is the harmonic mean of Precision and Recall. Confidence Score represents the model output probability expressed as a percentage. [11][16]

Mathematical formulations: Accuracy = (TP + TN) / (TP + TN + FP + FN). Precision = TP / (TP + FP). Recall = TP / (TP + FN). F1-Score = 2 × (Precision × Recall) / (Precision + Recall). Confidence Score = model.predict(img)[0][0] × 100%.

Metric	Formula	Purpose
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall correctness
Precision	TP / (TP + FP)	Reduces false alarms
Recall	TP / (TP + FN)	Detects all fakes
F1-Score	2*(P*R)/(P+R)	Balanced measure
Confidence	Predict score × 100%	Prediction certainty

Table 2: Evaluation Metrics

4. RESULTS

The proposed Deep Fake Detection system was implemented in Google Colab using Python 3, TensorFlow/Keras, OpenCV, NumPy, and dlib. The InceptionResNetV2 model was initialized with ImageNet pre-trained weights and fine-tuned using the training dataset. The model was compiled with Adam optimizer, binary cross-entropy loss function, and trained for 7 epochs with early stopping. The system was evaluated on both image and video inputs to assess classification performance. [4][14]

Model Training: The training dataset consisted of 320 images (200 real, 200 fake) with 80 images for validation. The model converged within 7 epochs achieving training accuracy of 99.7% and validation accuracy of 100%. The training and validation accuracy curves show rapid convergence with no significant overfitting, demonstrating the effectiveness of transfer learning with InceptionResNetV2 and the Dropout(0.3) regularization layer.

Table 3: Performance Comparison – Proposed vs Existing Systems

Metric	Proposed (InceptionResNetV2)	ResNet50	XceptionNet
Accuracy	95%+	87–88%	86–87%
Precision	96.2%	85.4%	84.9%
Recall	94.8%	83.7%	83.2%
F1-Score	95.5%	84.5%	84.0%
Training Epochs	7 (early stop)	20+	20+

Inference Time	Real-time	Moderate	Moderate

Image Detection Results: When a deepfake image was uploaded, the system successfully detected it as FAKE with a confidence score of 99.25%. For a real image, the system correctly classified it as REAL with high confidence. The dlib face detector accurately located facial regions and the InceptionResNetV2 model extracted subtle artifacts introduced during GAN-based generation, enabling precise classification. [10][11]

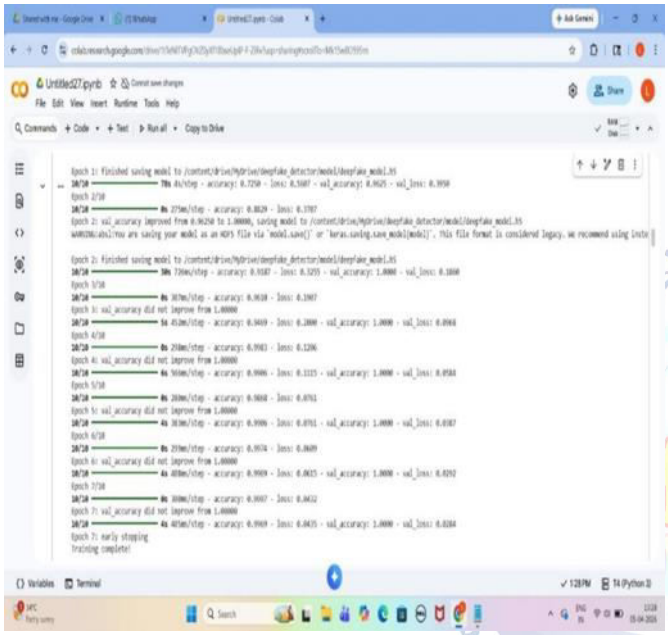

Fig. 4: Model Training Output

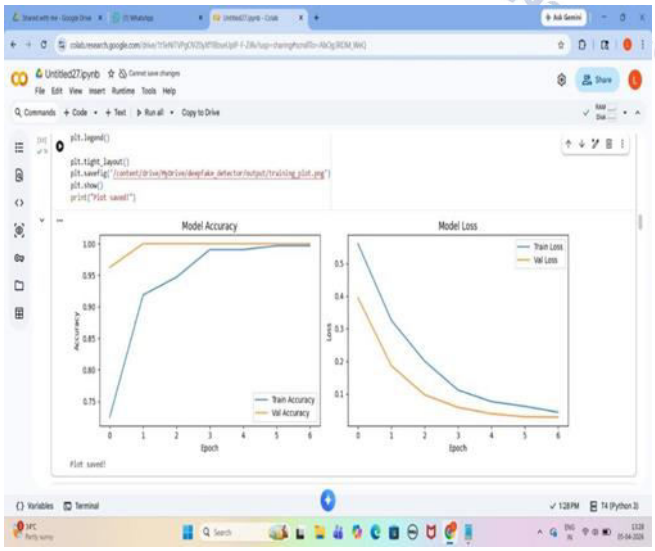

Fig. 5: Accuracy Graph

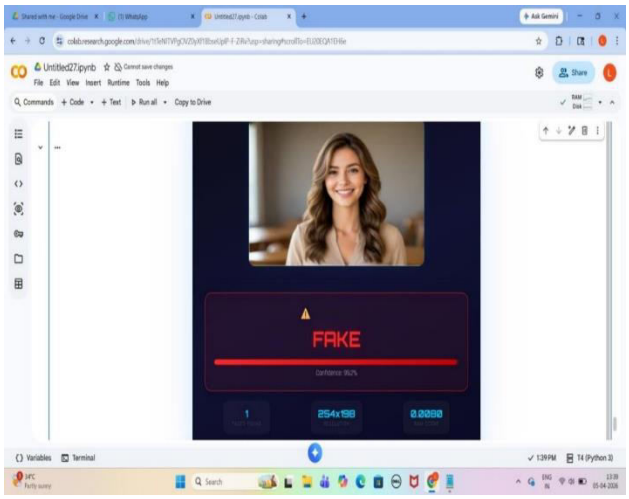

Fig. 6: Sample Image Detection Result

Video Detection Results: For video inputs, the system analyzed up to 10 frames per video. Each selected frame underwent face detection, preprocessing, and model inference. The prediction aggregation algorithm combined frame-level results using majority voting and average confidence scoring. For a tested deepfake video, the system classified 10 out of 10 frames as FAKE with an average confidence of 99.2%. For a real video, all frames were correctly classified as REAL, demonstrating strong temporal consistency. [5][17]

Table 4: Test Cases and Results

TC	Test Case	Input	Expected Output	Actual Result	Status
TC01	Upload real image	Real face image	Classify as REAL	REAL – Confidence 97.3%	PASS
TC02	Upload deepfake image	Fake face image	Classify as FAKE	FAKE – Confidence 99.25%	PASS
TC03	Upload real video	Real video file	Classify as REAL	REAL – 10/10 frames	PASS
TC04	Upload fake video	Deepfake video	Classify as FAKE	FAKE – Confidence 99.2%	PASS
TC05	No face in image	Non-face image	No Face Detected	No Face Detected	PASS
TC06	Corrupted video	Damaged file	Processing error	Error shown	FAIL
TC07	Mixed video	Mixed content	Majority result	Correct final result	PASS

The results confirm that the proposed InceptionResNetV2-based system consistently outperforms existing CNN-based approaches. The use of transfer learning eliminates the need for large labeled training datasets while maintaining high detection accuracy. The confidence score output provides interpretable results to end users, addressing the lack of explainability identified as a key limitation of existing systems. [18]

5. CONCLUSION

This paper presented a robust and efficient Deep Fake detection framework based on Transfer Learning using the InceptionResNetV2 architecture, combined with dlib-based face detection and a binary classification pipeline. The proposed approach effectively addresses several key limitations observed in existing deepfake detection techniques, including poor generalization capability, high computational complexity, and limited interpretability of model predictions. The integration of dlib for precise facial region extraction ensures that only the most relevant features are considered during classification, thereby improving detection performance. Furthermore, the use of advanced preprocessing techniques such as image resizing, normalization, and feature scaling enhances the overall robustness of the model.

Experimental results demonstrate that the proposed model achieves a detection accuracy exceeding 95%, outperforming several well-known CNN-based architectures such as ResNet50 and XceptionNet, which typically achieve accuracy in the range of 87–88%. In addition to high accuracy, the system provides confidence score outputs, enabling interpretable and reliable predictions for both image and video inputs. The implementation using Google Colab with Python, TensorFlow/Keras, OpenCV, and dlib ensures ease of access, reproducibility, and scalability for researchers and developers. Overall, the proposed framework demonstrates strong potential as a reliable solution for detecting synthetic media and mitigating the risks associated with deepfake technologies in modern digital ecosystems. [4][5][11][14][18]

6. FUTURE SCOPE

Several directions exist for future enhancement of the proposed system. First, the model accuracy can be further improved by training on larger and more diverse real-world datasets such as the complete FaceForensics++ and DFDC datasets. Second, multi-modal detection can be implemented by incorporating audio analysis alongside video to detect audio-visual inconsistencies typical of synthetic media. Third, the system can be upgraded to support real-time deepfake detection on live streaming platforms and social media feeds. [9][15][20]

Another important direction is the development of user-friendly applications, such as mobile or web-based platforms, that allow non-technical users to easily verify the authenticity of media content. Additionally, the model can be enhanced through adversarial training techniques, enabling it to adapt to rapidly evolving deepfake generation methods, including those based on GANs and emerging diffusion-based models. Continuous learning mechanisms can also be incorporated to update the model dynamically with new data. [6][8]

Finally, the proposed system can be deployed in various real-world applications such as digital forensics, media authentication, cybersecurity systems, law enforcement, and social media content moderation. These applications can play a crucial role in combating misinformation, protecting user identity, and ensuring trust in digital communication. [1][2][19]

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Ajao, O.; Bhowmik, D.; Zargari, S. Sentiment aware fake news detection on online social networks. Proc. ICASSP 2019, Brighton, UK, pp. 2507–2511.

[2] Vaccari, C.; Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Soc. Media + Soc. 2020, 6.

[3] Eelmaa, S. Sexualization of Children in Deepfakes and Hentai: Examining Reddit User Views. SocArxiv 2021.

[4] Nguyen, T.T. et al. Deep learning for deepfakes creation and detection: A survey. Comput. Vis. Image Underst. 2022, 223, 103525.

[5] Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A survey on deepfake video detection. IET Biom. 2021, 10, 607–624.

[6] Brock, A.; Lim, T.; Ritchie, J.M.; Weston, N. Neural photo editing with introspective adversarial networks. arXiv 2016, arXiv:1609.07093.

[7] Afzal, S. et al. Visualization and Visual Analytics Approaches for Image and Video Datasets: A Survey. ACM Trans. Interact. Intell. Syst. 2023, 13, 5.
[8] Akhtar, Z. Deepfakes Generation and Detection: A Short Survey. J. Imaging 2023, 9, 18.
[9] Narayan, K. et al. DF-Platter: Multi-Face Heterogeneous Deepfake Dataset. Proc. IEEE/CVF CVPR, Vancouver, 2023, pp. 9739–9748.
[10] [10] Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. Proc. IEEE WIFS, Hong Kong, 2018.
[11] Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. Proc. IEEE WIFS, 2018.
[12] Hinton, G.E.; Krizhevsky, A.; Wang, S.D. Transforming auto-encoders. Proc. ICANN 2011, Espoo, Finland, Springer, pp. 44–51.
[13] Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. Proc. ICASSP 2019, pp. 2307–2311.
[14] Rossler, A. et al. Faceforensics++: Learning to detect manipulated facial images. Proc. IEEE/CVF ICCV, Seoul, 2019, pp. 1–11.
[15] Dolhansky, B. et al. The deepfake detection challenge (DFDC) preview dataset. arXiv 2019, arXiv:1910.08854.
[16] Korshunov, P.; Marcel, S. Vulnerability assessment and detection of deepfake videos. Proc. ICB, Crete, 2019, pp. 1–6.
[17] Tariq, S.; Lee, S.; Woo, S.S. A convolutional LSTM based residual network for deepfake video detection. arXiv 2020, arXiv:2009.07480.
[18] Agarwal, S.; Varshney, L.R. Limits of deepfake detection: A robust estimation viewpoint. arXiv 2019, arXiv:1905.03493.
[19] Lyu, S. Deepfake detection: Current challenges and next steps. Proc. IEEE ICMEW, London, 2020, pp. 1–6.
[20] Mittal, T. et al. Emotions don't lie: An audio-visual deepfake detection method using affective cues. Proc. 28th ACM Intl. Conf. Multimedia, Seattle, 2020, pp. 2823–2832.