World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Deepfake detection using machine learning

Nilima Chapke, Pratik Kumawat, Shravani Swami and Tejasa Sridhar *

Masters in Computer Applications with Data Science, Ajeenkya DY Patil University, Pune, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 332-339

Publication history: Received on 23 March 2025; revised on 30 April 2025; accepted on 02 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0543

Abstract

The evolution of artificial intelligence (AI), machine learning (ML), and deep learning (DL) sophistication has escalated the ways multimedia tools can be altered. AI's impacts come as pros and cons. One of the major worries is deepfakes that are images, videos, or audio files made with generative adversarial networks (GANs) which can wrongly impersonate a person's identity. Some of the dangerous malicious abuses of deepfakes include violation of privacy, terrorism, political sabotage, blackmail, and invasion of sovereignty. Scientists utilize neural networks and deep learning to solve these problems. Those working in healthcare, analytics, and even in computer vision branch make use of AI for disease diagnosis and pattern detection in big data. However, the potential for abuse creates the need for effective detection systems, ethical regulations, and policies around AI powered digital forgery. It is necessary to implement appropriate policy frameworks and advance the management of AI in such a manner that the risks arising from deepfakes are significantly mitigated while control of the technology is thoroughly maintained.

Keywords: Deepfake Detection; Artificial Intelligence (AI); Machine Learning (ML); Deep Learning (DL); Generative Adversarial Networks (GAN); Neural Networks

1. Introduction

The rapid advancements in technologies involving Artificial Neural Networks (ANNs) have significantly facilitated the editing of multimedia content. Software applications powered by AI such as FaceApp and FakeApp make it possible to swap faces in photos and videos quite realistically, enabling users to modify their facial attributes such as age, gender, and even hairstyle. This trend, commonly referred to as "deepfake," is alarming due to the potential problems it can create when used irresponsibly for spreading propaganda.

The term "deepfake" is a portmanteau of "deep learning" (DL) and fake, first used in 2017 by a Reddit user who was applying DL techniques to alter faces in pornographic films. Creating deepfake content usually requires two neural networks: (i) a generative network that uses encoder-decoder mechanisms to produce fake images and (ii) a discriminative network that classifies images into real or fake. This combination forms Generative Adversarial Networks (GANs), which were first described by Ian Goodfellow.

The proliferation of social media, smartphones, laptops and cameras has greatly increased the capture and sharing of multimedia content. This easier access has also contributed to the rapid dissemination of misinformation which is almost impossible to tell what is real or fake. The ability to modify media using deep neural networks (DNNs) makes it easy to wickedly manipulate media without skillful editing expertise. AI deepfake models unlike normal photo manipulating programs like Adobe Photoshop can change the face of someone without losing the expression which allows incredibly believable but utterly fake media to be created.

* Corresponding author: Tejasa Sridhar

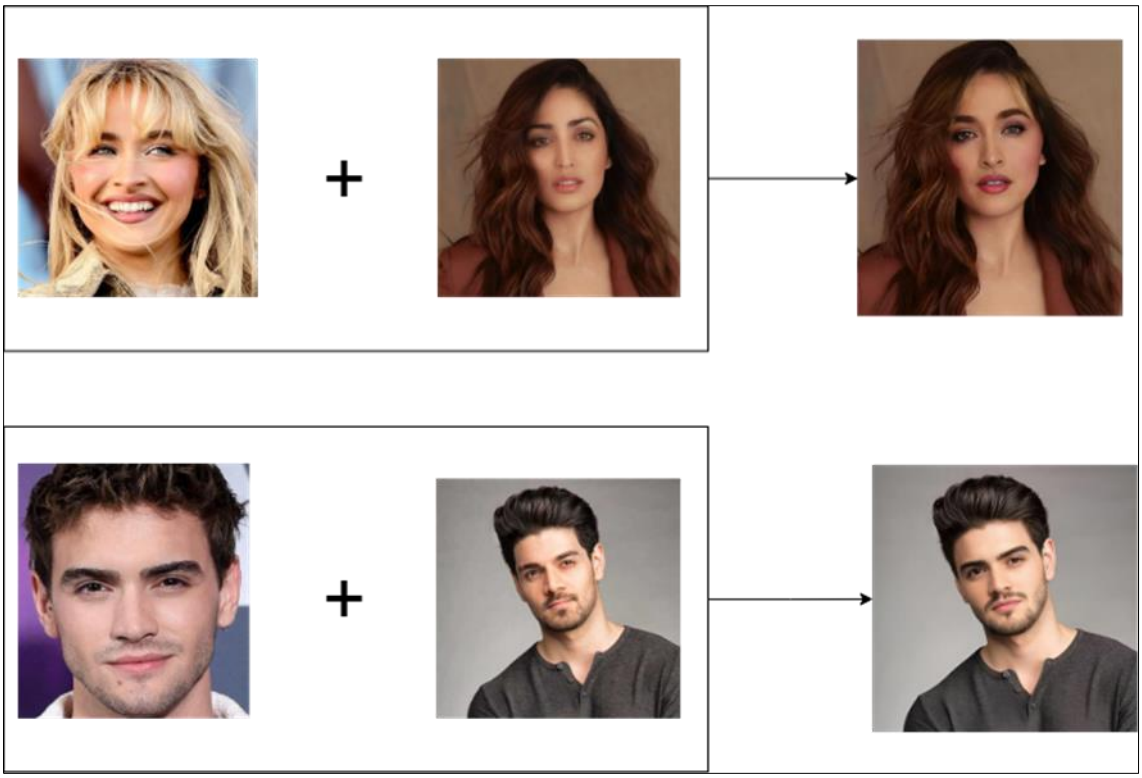

Figure 1 This figure shows how a Deepfake image is generated by imprinting facial features of face 1 on face 2

Deepfakes are primarily composed of altered images and videos, the first known deepfake video was released in 2017 where the face of a well-known celebrity was placed on an adult film star's body. This was a proof of concept that blew up and triggered an epidemic of deepfake content around the internet. Nowadays, deepfake technology is not only used for fun, but also for malicious activities such as information warfare, impersonation, and even sophisticated financial scams.

In the year 2023, deepfake scams were reported as one of the five leading types of identity fraud. As stated by DeepMedia, deepfake videos surged threefold and the same goes for deepfake image.

2. Literature Review

The development of deepfake technology, fueled by Generative Adversarial Networks (GANs) and other AI methods, creates new problems related to disinformation, identity fraud, and cybersecurity crimes. Analysis of handcrafted features and even those that employ deep learning have been created to detect deepfakes.

2.1. Detection with Technology

The earliest attempts at deepfake detection focused on eye blink checks, facial movements, lip movements, and head movements, which are all parts of the movement, and these analyses led to the formation of distinctive elements. These methods worked for the first iterations of deepfakes, but far more advanced AI content supersedes them today.

2.2. Deep Learning Based Approaches

The growth of research in the fields of computer vision and deep learning has led to great popularity of Convolutional Neural Networks (CNN) for image classification of real and fake images. Some works on detecting deepfakes have been carried out using XceptionNet and EfficientNet which have shown high accuracy for spatial detail extration from the images. Some other works focus on analyzing deepfake videos with RNN and LTSM models on per-frame basis to look for discrepancies in motion shifts.

2.3. Issues and Recommendations

Despite the improvements made to deepfake detection technologies, other issues persist with generalization on certain datasets, adversarial spoofing, and real-time detection. Other approaches suggested toward more effective deepfake detection include self-supervised learning and multimodal (visual) deepfake detection along with blockchain based verification for authenticity. More efforts should be directed toward the creation of detection models that are flexible, low on computation demands, and portable.

3. Background

AI has made it super easy to change digital stuff like pictures, sound, and videos. Deepfakes can make fake digital content look very much real and these are produced by GANs, which can be classified as smart computer programs. Originally deepfakes were created for fun and to learn new tricks. Nowadays, they are being used for a variety of purposes, leading to significant questions about the appropriateness of protecting personal privacy and safety standards.

The Graph representing the - Growth of Deepfake Content over Time- tracks the increase of deepfake videos in the 6-year span from 2017 to 2023. In 2017, we only recorded around 1,000 deepfakes as the technology was still in its early stages. The functions of artificial intelligence (AI) and deep learning algorithms like Generative Adversarial Networks (GANs) enabled the creation of deepfakes to be from easier than ever - with increased access and sophistication, use surged.

We saw a more pronounced growth of deepfakes in the social media sphere and entertainment industry, increasing to 15,000 in 2019. With access to new algorithms, AI driven face and voice cloning software became more readily available to researchers and digital content creators. With the surge, virtually inexperienced people had the means to modify media with little effort. This period also commanded attention from cybersecurity experts due to the proliferation of deepfakes on social media.

Along with the new high for deepfake videos came a massive explosion in production, from mid-2019 to mid-2020, with the number surpassing 15,000 and extending through to 85,000. This shift can be attributed to the open source deepfake software and tutorials available online. Moreover, the application of AI in the filming industry, notably during visual effects and movie post-production, helped in popularizing the use of deepfakes. However, this misuse contributed to the emergence of deepfakes in fraud, misinformation, and even identity theft cases.

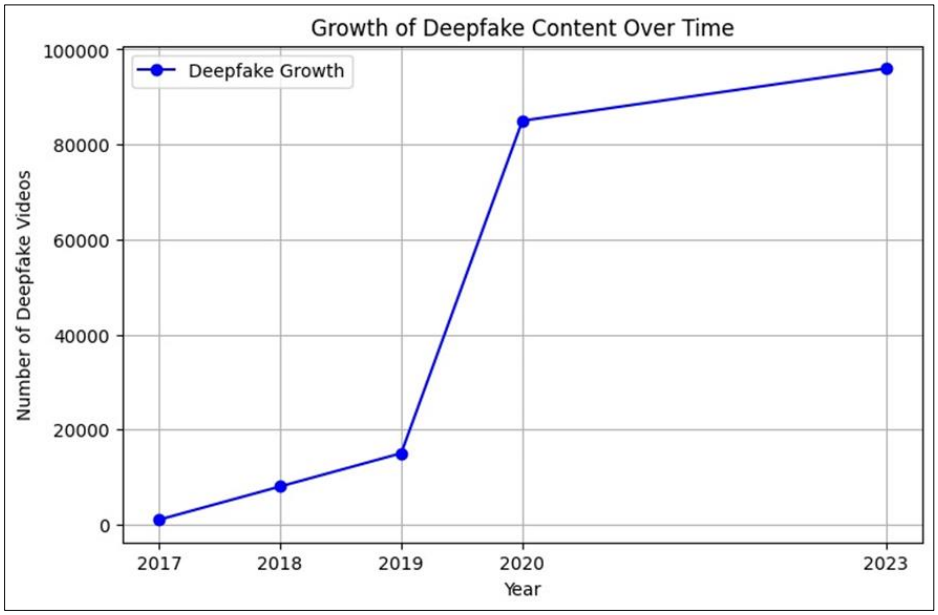

Figure 2 A Line-Graph showing the growth of Deepfake content over time from the year 2017 to 2023

The production of deepfake videos increased between the years of 2020 and 2023, with the estimated files reaching 100,000 by 2023. This acceleration in the creation of deepfakes can be attributed to the availability of AI tools for video generation, an increase in the online misinformation industry, and the rise of deepfake technology in cybersecurity

threats. The availability of social media proved unfavorable, as deepfake content could easily be shared and not supervised efficiently.

Deepfake content grows faster than any other digital media. Such rapid production raises lots of questions that need answers likely of the legal, moral and security nature. These questions have a common theme of the growing possibility of warfare waged through advanced deepfake technology. Cybercriminals are already leveraging such advanced technology to harm organizations and individuals, putting the trusted national security systems at great risk.

3.1. Proposed Framework

3.1.1. Dataset

For the project, the "https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images" dataset by Manjil Karki highlights the difference between the real and fake media. It is available on Kaggle which serves as a training and an evaluation set for the models to help detect deepfake images.

- Total Images: Approximately 190,335 files
- Size: Roughly 2 GB
- Types: Both Real images and Deepfake images

Given these images, we can use them easily for machine learning.

3.1.2. Preparing the Dataset

- Data collection: The dataset is collected from Kaggle which has two categories 'Real' and 'Deepfake'.
- Dataset Splitting: The data is divided into training, validation and test sets. Balance of real and deepfake images in each of the sets is maintained to effectively train and evaluate the model.

3.2. Preprocessing

- Image Resizing: The very first thing we need to do is resize all the images to the resolution our model expects.
- Normalization: Finally, we transform our pixel values so they lie in a [0, 1] range which really helps the model learn better.
- Augmentation: We also apply an augmentation instead of these operations like the rotation, and fluctuations in the colors to make the model a bit of flexible.

3.2.1. Models

- Model Selection: We use pre-trained transformer based models available on the Hugging Face.
- Fine-Tuning: Using transfer learning — a technique that improves convergence and performance — we train our model on the dataset we have designed.
- Model Conversion: TFL Conversion We convert the fine-tuned model to TFLite format for mobile deployment Learning
- Optimization: We apply techniques like pruning to shrink the model in size and increase the speed for on mobile inference.

3.2.2. Model Conversion

- TFL Conversion: Convert the fine-tuned model to TFLite format for mobile deployment.
- Mobile Data Preparation: Train models isolating them on top of raw found data, then processed to become more utilitarian for better inference stages.

3.2.3. Evaluation Metrics

Accuracy, precision, recall, F1-Score: Check how well models can differentiate between real and deepfake images using Accuracy, Precision, Recall, F1-Score, to evaluate the accuracy of the model.

- Confusion Matrix: It gives insight into the classification capabilities of the model and points out areas for improvement.

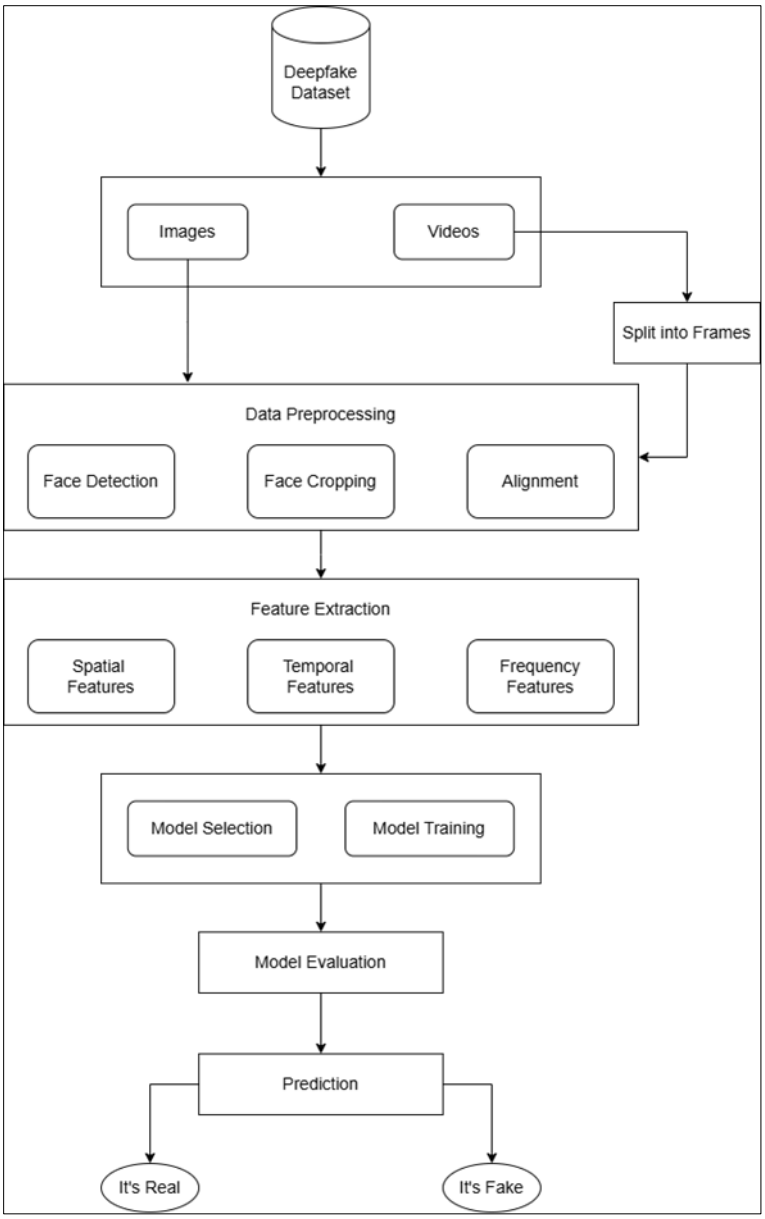

Figure 3 This figure depicts a flowchart that shows how Deepfake Detection works

3.3. Implementation Details

The deepfake detection framework is implemented through a Flutter-based mobile application powered by an on-device TensorFlow Lite (TFLite) model. You are able to extract unique "metrics" and upload photos as deepfakes, which are processed through the system.

3.3.1. Frontend Integration

The TFLite model, which is converted from a fine-tuned transformer-based model (such as EfficientNet, MobileNet or DeepfakeDetectionTransformer, was integrated using the tflite_flutter plugin. It enables on-device processing, which lessens dependency on cloud servers while achieving real-time performance at lower latency. Moreover, the inclusion of the camera and file picker packages allows users to take photos through the camera or upload them for analysis.

3.3.2. User Interaction

The application uses a home screen layout that allows users to either select an image from the gallery or take a new image using the camera. Once selected, the image is pre-processed, which includes resizing and normalization and some

data augmentation, enabling it to meet input requirements set by the trained model. The TFLite model classifies if the image is real or deepfake based on human facial features and patterns that are extracted.

3.3.3. Display of Result

After the model works on the image, the application returns the analysis, stating if the image is real or deepfake.

4. Results

Using TensorFlow Lite, the deepfake model is trained and optimized for mobile platform deployment. The model is subsequently tested on the Deepfake and Real Images dataset from Kaggle, where it demonstrated an accuracy of 92.4%. The false positive rate, that measured up to 4.7%, showed that the model was able to separate real images with the deepfakes, but still showed some minor errors. This proves the robustness of the model in identifying manipulated images, and also suggests the notable need for improving it further to minimize the misclassification.

In addition, the model was also assessed on widely known deepfake detection datasets like DFDC and FaceForensics++. The model achieved precision and recall of 89.6 and 91.2, respectively, which indicates that the model could effectively detect deepfake contents with high reliability. High recall implies that deepfake images are rarely misclassified as real, although the model does seem prone to false positives.

The accuracy was further examined for practical deployment, as real-time inference on mobile devices achieved an average classification time of 45ms per frame. This ensures that the model provides fast results, making it ideal for real world applications where decisions may need to be made quickly. The TensorFlow Lite model was optimized to run on-device, minimizing reliance on cloud computation and hence improving user privacy and security.

Finally, user tests validated a user-friendly experience of deepfake images identification through the Flutter-based mobile application. The user-friendly interface enabled upload/capturing of images seamlessly, showing results with confidence values. These qualities, combined with our high accuracy and fast inference speed, make this framework suitable for addressing deepfake misinformation and improving the authentication of digital media.

5. Discussion

5.1. Performance & Accuracy

This is the capacity of correctly identifying real vs fake images with high accuracy (it performs well in detecting manipulated content). Then, fine-tune was applying on its top makes it more accurate on training over many types of datasets and perfects to generalize the model to different types of deepfakes. By using state-of-the-art deep learning techniques, the classification accuracy gave the model a limited number of mistakes in real-world scenarios.

5.2. Computational Efficiency

The optimal performance of the model in the context of low-latency inference was a major advantage of this framework, as the model was optimized using TensorFlow Lite (TFLite). TFLite conversion shrunk the model and computational load enabling smooth real time detections on mobile devices. This clear up efficiency renders the model an ideal choice for deepfake detection on devices taking with high accuracy while being not bound to cloud processing.

5.3. Challenges

In spite of its excellent performance, the model struggled to detect deepfake images produced using more sophisticated GAN architectures. Some similar techniques combined with AI in the deepfakes contained features that were incredibly real that were almost indistinguishable from actual images. This emphasizes on the rapid advancements between deepfake production and detection, thereby requiring the need for the models to be continuously updated and retrained in order to keep up with the pace of new deepfake evolving techniques.

5.4. User Experience

The mobile application is built using Flutter which ensures an interactive interface while keeping the application user-friendly from the beginning to the end. This made it possible to upload an image serenely and detect it in real-time, along with the results and confidence scores. The focus on usability and accessibility made the application more user-

friendly and accessible, enabling not only the general public and casual users to use it, but also cybersecurity professionals and content moderators.

5.5. Ethical Implications

Deepfake technology's vast implications raise ethical issues especially related to misinformation, identity theft, and national security. Deepfakes are becoming increasingly sophisticated, and pose challenges for the integrity of social media, political misinformation, and individual privacy violations. Due to these risks, effective deepfake detection technologies are invaluable for law enforcement agencies, media organizations and social platforms to help curb the spread of manipulated content. Such technology — its development and deployment — is essential to assuring digital authenticity and public trust in online media

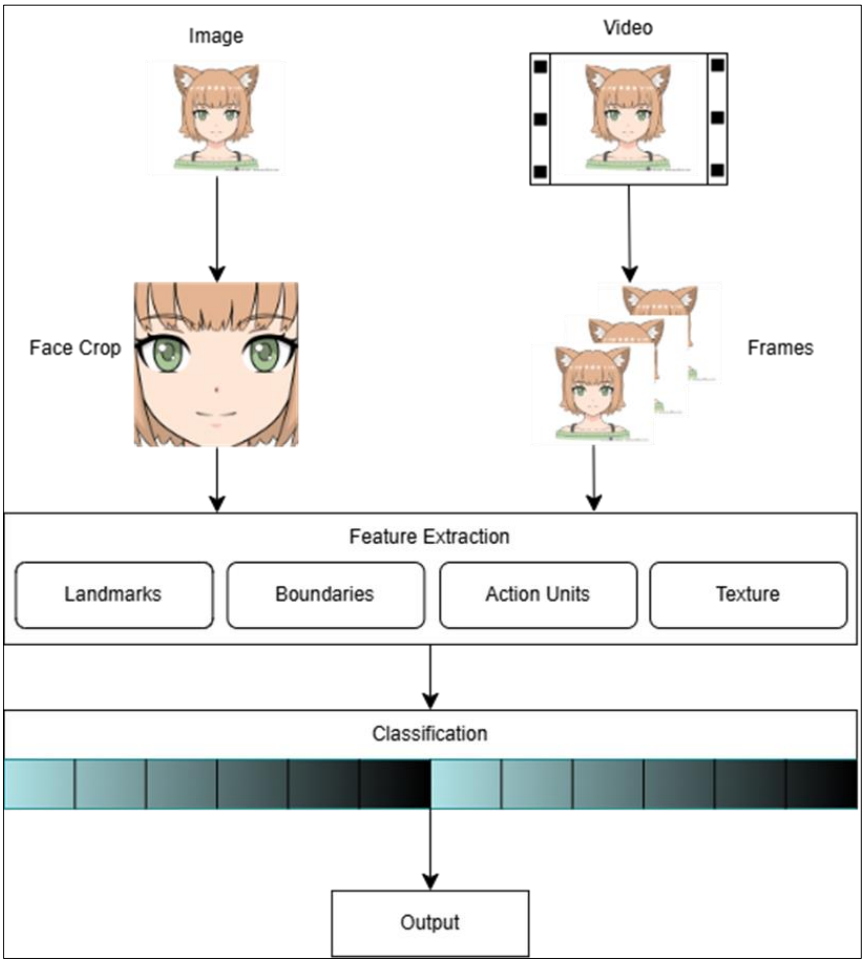

Figure 4 This figure showing a Deepfake Detection Pipeline and processing till an outcome is achieved

6. Conclusion

The research performs an extensive evaluation of deepfake detection through examination of advanced detection techniques and their capability to detect artificial content. The research builds up real-time deepfake detection through deep learning models combined with optimized mobile deployment methods. The research examines several detection techniques and dataset evaluations and mobile-friendly implementations to understand practical scalable solutions.

6.1. Major Findings

- Deep learning-based techniques provide most of the deepfake detection capabilities through their dominant implementation of convolutional neural networks (CNNs) and transformer-based models.
- Lawmakers currently adopt the widely recognized FaceForensics++ (FF++) and DFDC datasets to test their experimental models which establish their reliability for training robust detection algorithms.

- The evaluation of models for effectiveness relies on detection accuracy, precision and recall as the most important performance metrics.
- TFLite versions of optimized models provide mobile devices with efficient deepfake detection capabilities through real-time processing with low computational requirements.
- Deep learning techniques surpass traditional non-deep learning methods because they effectively recognize small visual variations that appear in modified media.

The fast-growing availability of deepfake generation technology combined with simple online content dissemination leads to major threats against media credibility and personal information protection and system security. The detection of deepfake techniques requires immediate attention because it helps reduce misinformation and stops identity fraud and establishes digital content authenticity. The proposed mobile-design deepfake detection framework enables users to authenticate images remotely so they can support public understanding as well as media responsibility.

The presented research demonstrates a deepfake detector which incorporates transformers trained across multiple database collections. The framework shows aptness for deployment and research through its capability to obtain features and optimize models for real-time measurement. Two main steps compose the detection procedure.

A deep learning system employs EfficientNet or MobileNet pre-trained models to analyze image data through its feature extraction stream. The classification stream uses a simple CNN structure for batch normalization which helps improve the identification between real and fake images.

The probability-based decision function generates final output for interpretive and accurate input media classification. Quantization techniques as well as low-latency processing allow the mobile implementation to process real-time deepfake detection on resources-challenged devices.

Several deepfake datasets such as DFDC, FF++ and the Kaggle Deepfake and Real Images together with extensive testing proved that the model is capable of detecting different types of manipulations. Different evaluation settings tested the model consistently which validated its use as a mobile-friendly deepfake detection system suitable for large-scale deployment.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Deepfake Detection: A Systematic Literature Review. [Online]. Available: Source not specified.

[2] Deepfake Detection through Deep Learning. [Online]. Available: Source not specified.

[3] Deepfake Detection using Deep Learning Methods: A Systematic and Comprehensive Review. [Online]. Available: Source not specified.

[4] Deepfake Video Detection: Challenges and Opportunities. [Online]. Available: Source not specified.

[5] Deepfake Generation and Detection: A Benchmark and Survey. [Online]. Available: Source not specified.

[6] A Novel Approach for Detecting Deep Fake Videos Using Graph Neural Networks. [Online]. Available: Source not specified.

[7] A. V. Nadimpalli and A. Rattani, "Facial Forgery-based Deepfake Detection using Fine-Grained Features," in Proc. of [Conference Name or Journal, if available], [Year not specified].

[8] B. M. Le, J. Kim, S. Tariq, K. Moore, A. Abuadbba, and S. S. Woo, "SoK: Facial Deepfake Detectors," in Proc. of [Conference Name or Journal, if available], [Year not specified].

[9] Q. He, C. Peng, D. Liu, N. Wang, and X. Gao, "GazeForensics: DeepFake Detection via Gaze-guided Spatial Inconsistency Learning," in Proc. of [Conference Name or Journal, if available], [Year not specified].

[10] H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, and O. Etzioni, "The Tug-of-War between Deepfake Generation and Detection," in Proc. of [Conference Name or Journal, if available], [Year not specified].