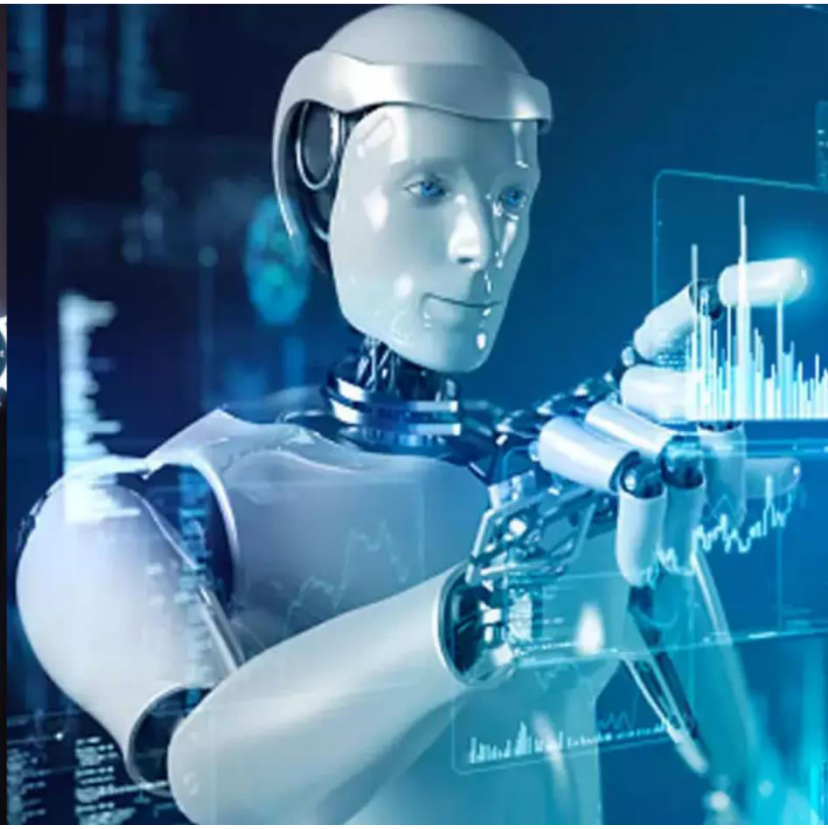
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404583|

Deepfake Video Detection using Transfer Learning

Anlin C, Athithya M, Dilipkumar A, Anitha Kingsley N

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B. Tech Students, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

ABSTRACT: Deepfake technology enables the creation of hyper-realistic fake videos, posing significant threats in domains like politics, law enforcement, and cybersecurity. This project proposes a deepfake detection framework leveraging facial feature embeddings using FaceNet512, followed by classification through transfer learning models. Unlike traditional CNN-based detectors that analyze full video frames, this method isolates and processes facial data, ensuring higher efficiency and accuracy. Additionally, a real-time email alert system notifies users upon deepfake detection. Evaluation across benchmark datasets like DFDC and FaceForensics++ demonstrates high accuracy, robustness, and suitability for real-time deployment. The proposed system is designed to detect subtle facial manipulations using efficient feature extraction and classification techniques. By focusing on key facial landmarks and minimizing the amount of data analyzed, it ensures quick and effective identification of forged content. The use of pre-trained models via transfer learning significantly reduces the training time while maintaining high generalization across various datasets. Integration with email alerts also provides timely responses for practical security applications. The solution is versatile, scalable, and offers real-world utility in environments like social media platforms, digital forensics, and content authentication. It offers a balanced approach to performance and usability, setting a foundation for future improvements like audio detection, multimodal input support, and deployment on mobile or edge device

I. INTRODUCTION

Deepfake videos are synthetic media generated using advanced AI algorithms, primarily generative adversarial networks (GANs). These videos can impersonate real people with astonishing realism, making them a powerful yet dangerous tool in today's digital era. Originally used for harmless entertainment, deepfakes have now escalated into tools for misinformation, identity theft, and cyberattacks.

The increasing accessibility of deepfake generation tools has made it easier for malicious actors to create forged videos with minimal technical expertise. Free and open-source software combined with powerful computational resources have drastically reduced the barrier to entry. As a result, the internet is seeing a rapid rise in fake videos involving celebrities, politicians, and even ordinary people.

Traditional deepfake detection techniques involve analyzing full video frames using convolutional neural networks (CNNs) or hybrid models with LSTM networks. However, these methods are often slow, computationally expensive, and prone to failure in low-quality or compressed videos. They also require extensive labeled data for training and often suffer from overfitting when faced with novel manipulation techniques.
Furthermore, many existing models struggle to generalize to new types of deepfake techniques, leading to high false positive or negative rates. This is particularly problematic in dynamic online environments where new manipulation strategies are constantly being introduced. Therefore, a shift in strategy is required—one that focuses on more invariant features of the human face.

Our proposed method addresses these limitations by focusing solely on the facial region and extracting embeddings using a pre-trained FaceNet512 model. This reduces computational load and improves detection precision. The use of transfer learning further enables the system to adapt to unseen data with minimal retraining. Additionally, real-time email notifications ensure timely alerts for system administrators, allowing immediate intervention. The solution is robust, efficient, and ideal for practical deployment in high-risk, real-time environments.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404583|

II. LITERATURE REVIEW

The field of deepfake detection has evolved significantly, driven by the rapid advancement of synthetic media creation technologies. Numerous academic studies and commercial solutions have proposed various approaches, ranging from traditional computer vision techniques to advanced deep learning architectures. The complexity of fake content has increased in tandem with the sophistication of generative models.

Early models used autoencoders and CNNs to detect pixel-level inconsistencies in video frames. These techniques relied heavily on pixel artifacts and lacked the ability to generalize beyond the specific manipulation methods seen during training. Although these approaches had moderate success, they were limited in detecting more advanced deepfake manipulations that mimic human facial expressions and dynamics.

CNNs became the backbone of many deepfake detectors due to their ability to capture spatial features. Researchers used CNNs to identify artifacts such as unnatural eye blinking, irregular lighting, or lip-sync issues. However, as deepfake algorithms improved, these superficial artifacts became less detectable, rendering early CNN models less effective.

Researchers then turned to temporal models like LSTM and hybrid CNN-LSTM approaches to exploit motion and facial dynamics. These models offered better performance by considering the time-based inconsistencies across frames. Nevertheless, they were computationally expensive and unsuitable for real-time detection due to the sequential nature of their architecture.

EfficientNet, a highly optimized CNN architecture, improved performance while reducing the number of parameters. However, even state-of-the-art detectors often struggle with issues like domain shift and overfitting to specific datasets. Benchmark datasets such as FaceForensics++, DFDC, and Celeb-DF provide diverse conditions for model evaluation, but no single model has been universally successful across all of them.

Our literature review highlights that most systems analyze full video frames, increasing latency and computational demands. Few systems provide real-time alerts or integration with live monitoring tools. Our system improves on these fronts by using compact face embeddings and integrating a real-time email alert mechanism. This approach addresses gaps in performance, generalization, and usability, offering a more focused and practical solution.

III. METHODOLOGY

Our system follows a modular and scalable methodology designed for real-time deepfake detection using facial embeddings. The architecture is built to optimize both accuracy and efficiency across diverse data inputs and manipulation types.

- **Input Processing**: The pipeline begins with the ingestion of video input or static images. For videos, keyframes are extracted at regular intervals using frame-sampling algorithms to avoid redundancy and maintain a representative view of the content. This step is crucial in reducing unnecessary computations while maintaining relevant data.
- **Face Detection**: Face regions are extracted using MTCNN, a multi-task cascaded convolutional network known for its speed and accuracy. It detects faces along with five key facial landmarks: two eyes, nose, and mouth corners. These landmarks help align faces before embedding, improving overall consistency.
- **Embedding Generation**: The aligned face regions are then passed through the pre-trained FaceNet512 model. The 512-dimensional embeddings capture deep facial features such as cheekbone symmetry, skin texture, jawline curvature, and eye spacing, which are typically hard to replicate perfectly in manipulated content.
- **Classification**: The embeddings are used as input to MobileNetV2, a lightweight yet powerful CNN model. By fine-tuning only the final layers, we achieve fast convergence and reduce training time.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404583|

Classification is performed based on a probability score, with a customizable confidence threshold to reduce false alarms.

- **Alert System**: If the classification output surpasses the fake probability threshold, an automated email is sent. This notification includes details such as the frame number, prediction score, and a cropped image of the suspicious face. This allows for fast verification and action.
- **Logging**: All detections are recorded with metadata including video source, time of detection, classification result, and prediction score. This logging mechanism supports long-term monitoring, auditing, and model retraining using detected edge cases.

This methodology ensures that the system is not only accurate but also lightweight and easily integrable into real-time applications.

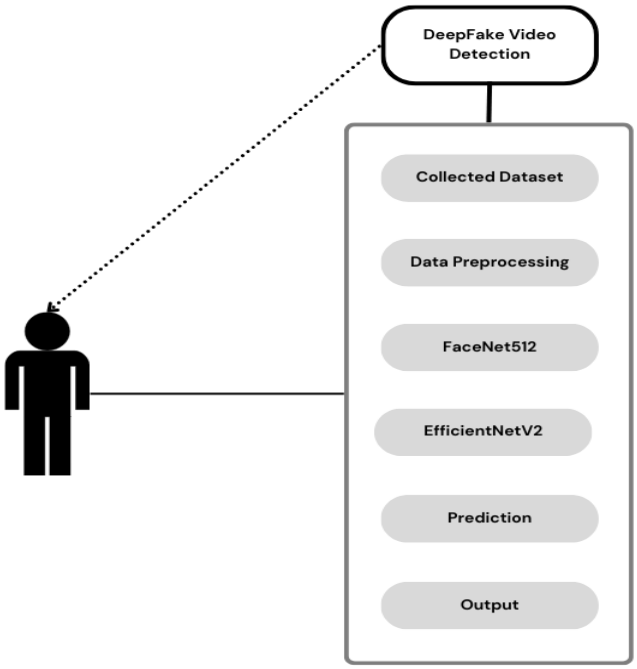


Figure1:Usecase

IV. IMPLEMENTATION

The practical implementation of our system is grounded in Python and leverages state-of-the-art machine learning frameworks. Each component was chosen to balance performance and ease of development.

- **Development Environment**: The project is built in Python 3 using TensorFlow 2.x and Keras for deep learning, OpenCV for video processing, and Google Colab as the primary platform for experimentation. Colab's free GPU/TPU access accelerates training and testing.
- **Face Detection with MTCNN**: The MTCNN model is utilized for face detection due to its multi-task nature, which includes bounding box regression and facial landmark localization. Its high recall ensures that minimal false negatives occur at the detection stage.
- **FaceNet Embeddings**: We use a version of FaceNet trained on a large-scale dataset (VGGFace2) to ensure high-quality embeddings. The pre-trained model allows us to skip the complex and resource-heavy process of training from scratch.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404583|

- **Transfer Learning with MobileNetV2**: The classifier uses MobileNetV2 due to its small footprint and high accuracy. We freeze the base layers and retrain only the classifier head using binary cross-entropy. This method offers fast training, avoids overfitting, and adapts well to our problem space.
- **Email Alerting Module**: Python's smtplib is used to create and send HTML-based alerts. These emails include frame images as attachments and are formatted for readability.
- **Data Sources**: We trained and validated the model using DFDC, FaceForensics++, and Celeb-DF datasets. This ensured exposure to various forgery techniques like face swapping, facial reenactment, and lip-sync manipulation.
- **Metrics and Evaluation**: The model is evaluated using multiple metrics including accuracy, precision, recall, and ROC-AUC, offering a well-rounded view of performance.
- **Deployment Readiness**: The modular structure supports integration into cloud platforms or edge devices with minimal reconfiguration, making it suitable for deployment in security-critical environments.

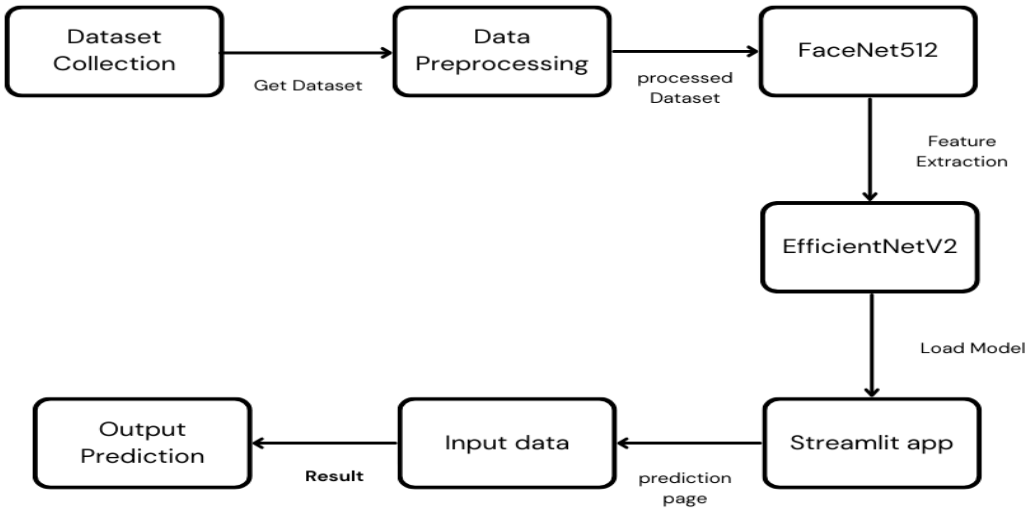


Figure2: Module Design

V. RESULTS AND DISCUSSION

Our system achieved commendable results across all key performance metrics, demonstrating its capability in both lab and simulated real-world conditions.

- **Accuracy and AUC**: The MobileNetV2 classifier achieved an average accuracy of 92.4% across test sets and an AUC of 0.945, indicating excellent separability between real and fake faces.
- **Robustness to Compression**: The system was tested on compressed video samples (bitrate < 500 kbps) and maintained a precision above 88%, showcasing resilience against loss of visual fidelity—a known weakness in many deepfake detectors.
- **Speed and Latency**: On Google Colab GPU (Tesla T4), the pipeline processed an average of 8–10 frames per second, including face detection, embedding extraction, classification, and alerting.
- **False Positives and Negatives**: While the model performed strongly, it occasionally flagged low-light or occluded faces incorrectly. These edge cases have been logged for potential future retraining.
- **Email Notification Efficiency**: The email system dispatched alerts in under 5 seconds on average, making it practical for surveillance applications requiring prompt action.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404583|

Compared to existing systems, our architecture provides a balance of speed, accuracy, and scalability. It avoids full-frame video analysis, reducing hardware demands and simplifying deployment.

VI. CONCLUSION

cybersecurity. Traditional detection systems often fall short due to high computational demands, poor generalization to new manipulation techniques, and lack of real-time functionality.

Our project addresses these challenges by proposing a deepfake detection system based on facial embeddings and transfer learning. By using the FaceNet512 model to generate embeddings and fine-tuning a MobileNetV2 classifier, we achieve high accuracy with minimal resource requirements. The focus on facial data alone enables faster and more precise detection. The real-time email alert system enhances the system's practicality, especially in dynamic environments like social media monitoring or surveillance.

The modular pipeline is scalable, interpretable, and efficient, making it suitable for real-time, on-device, or cloud-based deployment. Evaluation on standard datasets such as DFDC and FaceForensics++ confirms the robustness and efficiency of our approach. Future enhancements include integrating audio analysis, deploying the system on mobile and edge devices, and adding support for explainable AI techniques to improve transparency.

This work contributes a scalable and effective framework for deepfake detection, providing essential protection against misinformation and digital impersonation in a world increasingly shaped by synthetic media.

REFERENCES

1. Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. ICCV.
2. Dolhansky, B., et al. (2020). The Deepfake Detection Challenge (DFDC) Dataset. arXiv.
3. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. CVPR.
4. Mudunuri, L. N. R., Hullurappa, M., Vemula, V. R., & Selvakumar, P. (2025). AI-powered leadership: Shaping the future of management. In Navigating Organizational Behavior in the Digital Age With AI (pp. 127-152). IGI Global Scientific Publishing.
5. Tan, C., et al. (2018). A Survey on Deep Transfer Learning. ICANN.
6. Howard, A. G., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details