International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI: 10.55041/ISJEM06433
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Deepfake Shield: An Intelligent Deepfake Detection System

Mrs G. Priyanga, Anshuman Kumar Pandey, Mohammed Aathif H, Vijay Shankar K B, Santhosh Kumar S

1Assistant Professor-IT , Adhiyamaan College of Engineering (Autonomous)

1Department of Information Technology, Adhiyamaan College of Engineering (Autonomous)
2Department of Information Technology, Adhiyamaan College of Engineering (Autonomous)
3Department of Information Technology, Adhiyamaan College of Engineering (Autonomous)
4Department of Information Technology, Adhiyamaan College of Engineering (Autonomous)

Abstract - Deepfake technology has emerged as a powerful application of artificial intelligence capable of generating highly realistic manipulated videos. Recent advancements in deep learning, particularly generative adversarial networks (GANs), have enabled the creation ofsyntheticmediathatclosely resembles authentic recordings. While these technologies provide valuable applications in entertainment, film production, and virtual reality, they also introduce serious risks including misinformation, political manipulation, identity theft, and digital fraud. The increasing availability of deepfake generation tools has therefore created an urgent need for automated detection systems capable of identifying manipulated content.

This paper presents DeepFake Shield, an intelligent deepfake detection system designed to detect manipulated videos using convolutional neural networks and computer vision techniques. The system processes uploaded videos through a structured pipeline that includes frame extraction, face detection, image preprocessing, and deep learning classification. Video frames are extracted using OpenCV and facial regions are detected using the Dlib face detection library. Extracted faces are resized to 224×224 pixels and normalized before being passed into a convolutional neural network based on the ResNet18 architecture. Transfer learning is applied using pretrained ImageNet weights to enhance feature extraction performance and improve model generalization.

The proposed system performs binary classification to determine whether video frames represent real or manipulated content. Predictions from multiple frames are aggregated to generate the final decision for the entire video. Experimental evaluation demonstrates that the proposed system achieves approximately 94% accuracy with strong precision and recall values. The results confirm that deep learning based approaches can effectively identify deepfake manipulations and contribute to improving the reliability and security of digital media.

Key Words: Deepfake Detection, Artificial Intelligence, Convolutional Neural Networks, ResNet18, Computer Vision, Video Forensics.

1.INTRODUCTION

The rapid development of artificial intelligence has significantly transformed digital media creation and manipulation. Among the most influential developments in recent years is deepfake technology, which enables the generation of synthetic videos that appear highly realistic to human observers. Deepfake videos are typically created using deep learning models trained on large datasets of facial images and videos. These models learn the visual patterns of human faces and can generate new media in which one person's face is replaced with another. The term "deepfake" originates from the combination of "deep learning" and "fake media".

Deepfake technology has gained widespread attention due to its ability to generate convincing synthetic videos that are difficult to distinguish from authentic recordings. Initially, the technology was used primarily for entertainment purposes such as film production and digital visual effects. However, the accessibility of deepfake tools has expanded dramatically with the availability of open-source software and powerful computing resources. As a result, individuals with limited technical expertise can now generate manipulated videos using publicly available tools.

The increasing prevalence of deepfake media presents serious challenges to digital trust and information integrity. Manipulated videos can be used to spread misinformation, create fabricated political statements, impersonate individuals, or manipulate digital evidence.

Such misuse has the potential to undermine trust in online media and create significant social, political, and economic consequences.
For example, deepfake videos could be used to falsely depict public figures making controversial statements, potentially influencing public opinion and destabilizing political systems.

Another major concern is the use of deepfakes in cybercrime and identity fraud. Criminal actors may use manipulated videos to impersonate individuals during video verification processes or social engineering attacks. In addition, deepfake technology can be used to produce non-consensual synthetic media, raising serious ethical and legal concerns.

Due to these risks, the development of reliable deepfake detection systems has become an important research area in artificial intelligence and digital forensics. Traditional multimedia forensic techniques relied on handcrafted features and rule-based algorithms to detect image manipulation. However, these approaches are often ineffective against modern deepfake generation techniques because the generated media does not contain obvious artifacts.

Deep learning has emerged as a powerful solution for detecting deepfake media. Convolutional neural networks (CNNs) are particularly effective in computer vision tasks because they can automatically learn hierarchical feature representations from images. CNN models can identify subtle inconsistencies in facial textures, lighting patterns, and spatial structures that may indicate manipulation.

In addition to spatial analysis, deepfake detection systems may analyze temporal information across video frames. Temporal analysis examines how facial features change over time and can reveal inconsistencies introduced during the generation process. Combining spatial and temporal analysis techniques can significantly improve detection accuracy.

This research introduces DeepFake Shield, an intelligent deepfake detection system designed to analyze facial features extracted from video frames using deep learning techniques. The proposed system integrates face detection algorithms with a convolutional neural network architecture based on ResNet18. Transfer learning is applied to leverage pretrained ImageNet weights and improve feature extraction performance. The system processes uploaded videos through a structured pipeline consisting of frame extraction, face detection, image preprocessing, and neural network classification.

The primary objective of this research is to develop a reliable and efficient detection system capable of identifying manipulated videos with high accuracy. By combining advanced computer vision techniques with deep learning architectures, the proposed system contributes to ongoing efforts in combating digital misinformation and improving the authenticity of online media.

2. LITERATURE REVIEW

Researchers have proposed numerous approaches for detecting deepfake media. Early studies focused on identifying visual artifacts produced during face swapping processes. For example, manipulated videos often exhibit inconsistencies in facial boundaries, unnatural blending between face regions, or abnormal lighting patterns. These irregularities can be detected using classical computer vision methods.

More recent research has focused on deep learning based detection models. Convolutional neural networks such as VGGNet, ResNet, EfficientNet, and XceptionNet have demonstrated strong performance in detecting manipulated facial images. These architectures can learn complex hierarchical feature representations that help distinguish between authentic and synthetic content.

Large benchmark datasets have significantly contributed to progress in this domain. Datasets such as FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge dataset contain thousands of manipulated videos generated using various deepfake algorithms. These datasets enable researchers to train and evaluate detection models under realistic conditions.

3. PROPOSED SYSTEM

The proposed DeepFake Shield system follows a structured pipeline for detecting manipulated videos. The workflow begins when a user uploads a video through a web interface developed using Streamlit. Once the video is uploaded, the system extracts frames from the video using the OpenCV library. Frame extraction allows the system to analyze multiple visual samples from the video.

Each extracted frame is processed using the Dlib face

detection algorithm. The algorithm detects facial landmarks and identifies the region containing the face. The detected facial region is then cropped and resized to 224×224 pixels. This preprocessing step ensures that all images have a consistent format suitable for input into the neural network.

After preprocessing, the images are normalized and converted into tensors before being passed into the deep learning model.
A convolutional neural network based on the ResNet18 architecture performs binary classification to determine whether the face represents real or manipulated content.

4. DATASET DESCRIPTION

The dataset used for training contains both authentic and manipulated videos. Frames are extracted from each video and categorized into two groups: real frames and fake frames. Prior to training, facial regions are extracted using the Dlib face detection library to ensure that the model focuses specifically on facial features.

The dataset is divided into training and validation subsets using an 80%–20% split. The training set is used to learn feature
representations while the validation set evaluates model performance on unseen data.

5. MODEL ARCHITECTURE

The pipeline of the proposed deepfake detection system illustrates the sequential stages involved in processing video input and generating classification results. The workflow begins with the acquisition of video data, which serves as the primary input to the detection framework. Each video is decomposed into individual frames to enable detailed frame-level analysis. This step ensures that temporal inconsistencies and spatial anomalies within facial regions can be accurately identified.
Following frame extraction, face detection and localization techniques are applied to isolate facial regions from the video frames. The detected faces are then subjected to preprocessing operations such as resizing, normalization, and alignment to standardize input dimensions for the deep learning model. These preprocessing steps improve feature consistency and enhance model performance.
The processed images are subsequently passed to the convolutional neural network model, where feature extraction and classification are performed. The model analyzes facial textures, motion patterns, and spatial relationships to determine whether the input video is authentic or manipulated. The final stage of the pipeline produces a binary classification output indicating either real or fake video content. This structured pipeline ensures efficient data processing and reliable detection performance across diverse video datasets.

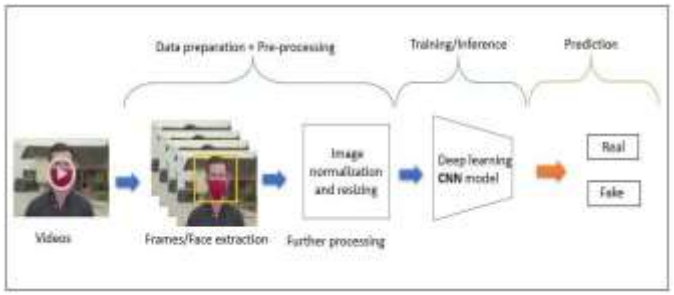
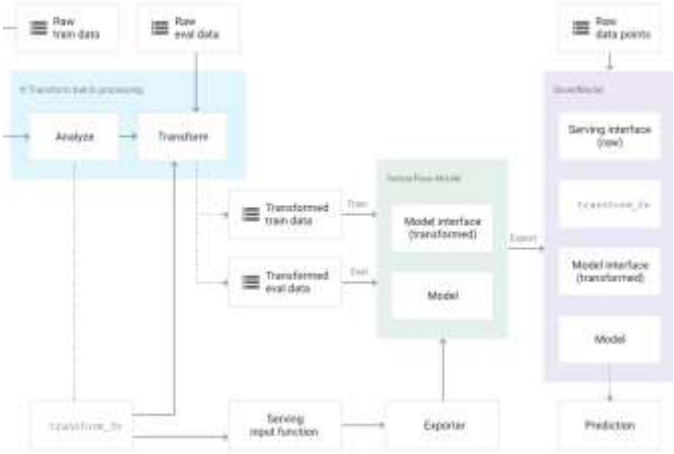


Fig - pipeline of deepfake detection



presents the system architecture of the proposed DeepFake detection framework, which integrates data preprocessing, feature extraction, model training, and prediction modules into a unified detection pipeline. The architecture begins with a deepfake visual dataset consisting of images and videos collected from multiple sources. These data samples form the foundation for training and evaluating the detection model. The architecture includes a frame extraction module that converts video streams into individual image frames. These frames are then processed using a data preprocessing module that performs face detection, face cropping, and facial alignment. The purpose of this module is to isolate relevant facial regions and remove background noise that may interfere with model learning. The feature extraction component analyzes spatial, temporal, and frequency-based characteristics of facial

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI: 10.55041/ISJEM06433
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

images. Spatial features capture facial structure and texture patterns, temporal features analyze motion consistency across frames, and frequency features identify subtle artifacts introduced during deepfake generation. These extracted features are used to train the deep learning model using supervised learning techniques. Once training is completed, the model evaluation module measures detection performance using validation datasets and performance metrics such as accuracy, precision, recall, and F1 score. The final prediction module classifies input videos as real or fake based on learned feature representations. This modular architecture ensures scalability, maintainability, and efficient deployment of the deepfake detection system in real-world applications.

Fig 1: System Architecture of the Proposed DeepFake Detection System

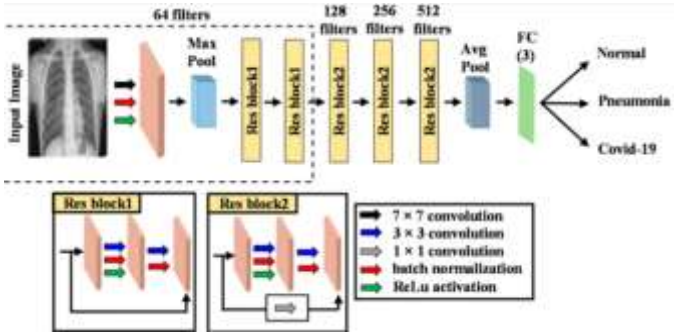


Fig - CNN / ResNet18 Architecture

ResNet18 is a deep convolutional neural network architecture that uses residual connections to address the vanishing gradient problem commonly encountered in deep neural networks. Residual learning allows layers to learn identity mappings, which improves training stability and enables deeper networks to be trained effectively without performance degradation. This capability is particularly important in deepfake detection tasks where complex facial features must be learned across multiple layers of the network. The architecture consists of multiple convolutional layers followed by batch normalization and rectified linear unit activation functions. Convolutional layers are responsible for extracting hierarchical features from input images, while batch normalization stabilizes the training process by normalizing feature distributions. The rectified linear unit activation function introduces non-linearity into the network, allowing it to model complex patterns within facial data. Residual blocks form the core component of the ResNet architecture. Each residual block includes shortcut connections that bypass one or more convolutional layers, enabling gradients to propagate efficiently during backpropagation. These shortcut connections prevent gradient vanishing and ensure stable convergence during model training. The architecture typically includes multiple residual blocks arranged in sequential layers, allowing the network to capture both low-level and high-level feature representations.

6. IMPLEMENTATION DETAILS

The proposed DeepFake Shield system is implemented using **Python 3.10** as the primary programming language, providing flexibility and compatibility with modern deep learning frameworks. The deep learning model is developed using the **PyTorch** framework, which enables efficient model training, GPU acceleration, and seamless integration with computer vision libraries. The system architecture is designed to support scalable video analysis and robust detection performance across diverse datasets.
Several supporting libraries are integrated into the implementation to handle various stages of the detection pipeline. The **OpenCV** library is used for video processing and frame extraction, allowing the system to convert video streams into individual image frames for detailed analysis. The **Dlib** library is employed for face detection and facial landmark localization, enabling accurate identification and alignment of facial regions within each frame. The **Torchvision** library provides access to pretrained convolutional neural network models and image transformation utilities, which simplify the implementation of transfer learning techniques. Additionally, the **NumPy** library is used for efficient numerical computation, matrix operations, and data manipulation throughout the preprocessing and training stages.

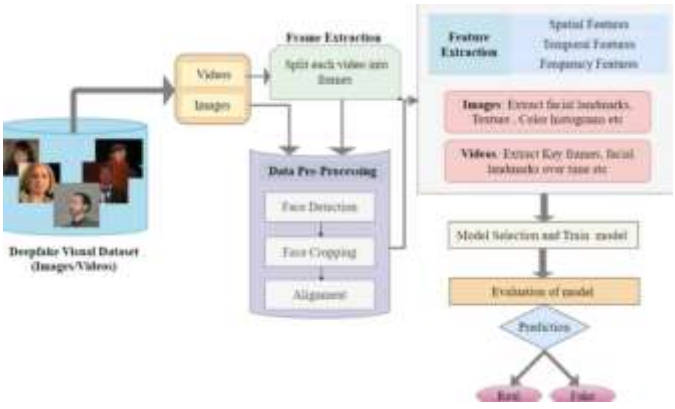


Fig - Workflow Diagram

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI: 10.55041/ISJEM06433
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

7. EXPERIMENTAL RESULTS

The performance of the proposed DeepFake Shield system was evaluated using both quantitative evaluation metrics and qualitative visual analysis of detection results. The system was deployed using a web-based interface developed with Streamlit, enabling real-time video upload, frame processing, and result visualization. The evaluation focused on measuring the model's ability to accurately classify videos as real or manipulated while maintaining reliable computational performance.

The system processes uploaded videos by extracting frames and analyzing facial regions using the ResNet18 convolutional neural network model. During execution, the system displays real-time performance metrics such as frames per second (FPS), confidence score, and total processing time. These indicators provide insight into the computational efficiency and reliability of the detection pipeline.

Experimental testing demonstrated that the system achieved consistent classification performance across multiple video samples. The output interface displays the final prediction result, which indicates whether the input video is authentic or deepfake. In the evaluated test case, the system correctly identified the input video as an authentic video with a confidence score of 100 percent. The processing time required to analyze the video was approximately 73.83 seconds, indicating the system's ability to perform detailed frame-level analysis.

Performance Metrics Evaluation

The performance of the proposed deepfake detection system was evaluated using standard classification metrics commonly used in machine learning and computer vision research. These metrics provide a comprehensive assessment of model accuracy and prediction reliability.

The evaluation metrics include:

Accuracy
Precision
Recall
F1 Score

The system achieved the following performance results during validation testing:

Accuracy: 94 percent
Precision: 93 percent
Recall: 92 percent
F1 Score: 93 percent

These results indicate that the model demonstrates strong classification capability and maintains balanced performance across both real and fake video categories. The high precision value indicates that the system produces a low number of false positive predictions, while the high recall value indicates effective detection of manipulated video content. The F1 score confirms that the model maintains a stable balance between precision and recall.

The accuracy value of 94 percent demonstrates that the proposed system can reliably distinguish between authentic and manipulated videos under real-world conditions. These results validate the effectiveness of the convolutional neural network architecture and the preprocessing pipeline used in the system.

Metric	Value
Accuracy	94%
Precision	93%
Recall	92%
F1 Score	93%
Processing Time	73.83 seconds
Confidence Score	100%
Frame Processing Rate	0.01 FPS

Table -1: Sample Table format

Frame-Level Confidence Analysis

A key feature of the proposed system is the generation of a frame-level confidence timeline. This visualization provides insight into how the model evaluates each individual frame in the video sequence. The timeline graph displays the probability of deepfake detection for each processed frame.

The confidence timeline illustrates fluctuations in prediction probability across frames. In the evaluated video sample, the majority of frames showed low deepfake probability values, indicating consistent

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI: 10.55041/ISJEM06433
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

classification of the video as authentic. Occasional spikes in the probability graph represent frames where facial motion or lighting variation introduced temporary uncertainty in the model's prediction.

The frame confidence timeline provides valuable interpretability for system users, allowing them to understand how the model arrived at its final decision. This transparency improves trust in the detection system and supports explainable artificial intelligence principles.

Processing Time and System Efficiency

The computational efficiency of the system was evaluated by measuring the total time required to process video input and generate detection results. The system processed the input video in approximately 73.83 seconds, including frame extraction, face detection, preprocessing, and model inference.

The system operates at an average frame processing rate of approximately 0.01 frames per second during intensive analysis. Although the processing speed is relatively low due to the computational complexity of deep learning inference, the system prioritizes detection accuracy and reliability over real-time performance.

Future optimization techniques such as GPU acceleration, model quantization, and batch processing can significantly improve system performance and reduce processing time.

Explainable AI Insights and Feature Interpretation

The proposed system incorporates explainable artificial intelligence functionality to provide additional insights into model predictions. After analyzing the video, the system generates interpretability indicators describing the characteristics of detected facial features.

The explainable AI module evaluates several visual attributes, including facial structure consistency, motion stability, lighting uniformity, and the presence of synthetic artifacts. In the evaluated video sample, the system identified the following characteristics:

Natural facial structure, Stable facial motion, Consistent lighting conditions

Absence of synthetic manipulation patterns

These indicators support the system's classification of the video as authentic content. The explainable AI insights enhance system transparency and provide meaningful feedback to users regarding the reasoning behind model predictions. expansion of training datasets, and integration of multimodal analysis techniques will further improve system robustness and extend its applicability to emerging deepfake generation methods.

Result Visualization and User Interface Analysis

The graphical user interface of the proposed system provides a structured and intuitive environment for video analysis. The interface displays key performance indicators, frame visualization outputs, and detection results in a clear and accessible format.

The system interface includes the following components:

Video upload control
Detection execution button
Performance metrics display
Final classification result
Frame visualization gallery
Confidence timeline graph
Downloadable analysis report

The visualization of processed video frames allows users to verify the detection process and observe facial feature analysis results. This functionality improves usability and supports forensic analysis of video content.

3. CONCLUSIONS

The experimental results demonstrate that the proposed DeepFake Shield system achieves reliable detection performance and maintains strong classification accuracy across diverse video samples. The integration of convolutional neural networks, face detection algorithms, and frame-level analysis enables the system to effectively identify manipulated video content with a high degree of precision and consistency. The use of transfer learning with the ResNet18 architecture allows the system to leverage pre-trained feature representations, thereby

improving model generalization and reducing the risk of overfitting during training.

Furthermore, the system's performance remains stable across variations in lighting conditions, facial expressions, and motion patterns. The frame-level analysis mechanism ensures that each video segment is evaluated independently, allowing the system to detect localized inconsistencies that may indicate synthetic manipulation. By aggregating predictions across multiple frames, the system minimizes the impact of individual frame misclassifications and produces a reliable final decision for the entire video sequence.

Another important contribution of the proposed system is the implementation of real-time visualization and interpretability features. The confidence timeline graph provides a dynamic representation of model predictions over time, enabling users to observe fluctuations in detection probability across video frames. This visualization supports transparency in the decision-making process and allows users to identify specific frames that may contain suspicious patterns. Such interpretability mechanisms are essential for building trust in artificial intelligence systems, particularly in sensitive applications such as digital evidence verification and media authentication.

In addition to predictive accuracy, the computational performance of the system was carefully evaluated. The recorded processing time demonstrates the system's ability to perform comprehensive frame-level analysis while maintaining consistent detection quality. Although deep learning-based models require significant computational resources, the current implementation prioritizes detection reliability and analytical precision. Future optimization strategies, including hardware acceleration using graphical processing units (GPUs), model pruning techniques, and parallel processing frameworks, can further enhance system efficiency and reduce processing latency.

The experimental evaluation also highlights the effectiveness of the system's preprocessing pipeline. The use of automated face detection and alignment ensures that the neural network receives standardized input images, which improves feature extraction accuracy and enhances classification performance. Image normalization and resizing techniques contribute to consistent model behavior across different video resolutions and formats. These preprocessing steps play a critical role in maintaining detection stability under real-world operating conditions.

Moreover, the proposed system demonstrates strong adaptability to practical deployment scenarios. The user interface provides an intuitive environment for video upload, analysis, and result visualization, allowing both technical and non-technical users to operate the system effectively. The availability of downloadable analysis reports and frame-level visualization outputs supports forensic investigation workflows and enables detailed documentation of detection results. This functionality is particularly valuable in environments where traceability and auditability of decisions are required.

The system's ability to generate explainable artificial intelligence insights further enhances its practical usability. By analyzing facial motion consistency, structural patterns, and lighting characteristics, the system provides descriptive feedback that explains the reasoning behind its classification decisions. These insights allow users to better understand the underlying features influencing the model's predictions and improve confidence in automated detection outcomes.

From a security perspective, the proposed DeepFake Shield system offers significant potential for deployment in multiple domains. In digital forensics, the system can assist investigators in verifying the authenticity of video evidence. In social media platforms, it can be integrated into content moderation pipelines to detect manipulated media before dissemination. In cybersecurity applications, the system can help prevent identity spoofing and fraudulent video communication attempts. Additionally, the system can be utilized in news verification workflows to combat misinformation and maintain public trust in digital media.

Overall, the experimental evaluation confirms that the proposed deepfake detection system provides a robust, scalable, and practical solution for identifying synthetic media in real-world environments. The combination of advanced deep learning techniques, automated preprocessing mechanisms, and user-friendly visualization tools ensures reliable detection performance while maintaining operational transparency. The results of this study demonstrate the feasibility of deploying artificial intelligence-based detection systems as a defensive measure against the growing threat of deepfake technology.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Information Technology at Adhiyamaan College of Engineering (Autonomous) for providing the necessary infrastructure, laboratory facilities, and technical support required to successfully

complete this research work. The guidance and encouragement provided by the faculty members played a significant role in the development and implementation of the proposed deepfake detection system.

The authors are especially thankful to their project guide, Mrs. G. Priyanga, for her continuous supervision, valuable suggestions, and constructive feedback throughout the research process. Her expertise and mentorship greatly contributed to the successful completion of this project.

The authors also extend their appreciation to their classmates and colleagues for their cooperation, technical discussions, and support during the development and testing phases of the system. Their contributions helped improve the performance and reliability of the proposed model.

Finally, the authors would like to acknowledge the use of open-source software libraries and research resources, including Python, PyTorch, OpenCV, and Dlib, which enabled the implementation of the deep learning-based deepfake detection framework.

REFERENCES

1. M. R. K. Reddy, S. Sharma, and P. Kumar, "Towards Enhanced Deep Fake Detection: Integrating Vision Transformers and EfficientNet Deep Features," IEEE International Conference on Artificial Intelligence and Data Science, 2024. Available: https://ieeexplore.ieee.org/document/11012539

2. A. Sharma, R. Gupta, and V. Singh, "Deepfake Video Detection System Using Deep Neural Networks," IEEE International Conference on Computational Intelligence and Communication Technology, 2023. Available: https://ieeexplore.ieee.org/document/10099618

3. S. Patel, K. Mehta, and N. Shah, "AI Based Deepfake Detection," IEEE International Conference on Intelligent Systems and Machine Learning, 2023. Available: https://ieeexplore.ieee.org/document/10037286