Deep Fake Detection in Images and Videos Using Spatial and Temporal Analysis

K. Subha[1], Aduri Sai Prateik[2], Gudla Sai Manikanta[2] and Nistala Sri Harsha[2]
[1]*Department of IT, St. Joseph's College of Engineering, OMR, Chennai, Tamil Nadu, India*
[2]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India*

Keywords: LSTM, RNN, CNN, GAN, ResNext.

Abstract: The rapid progress of deep fake technology has created opportunities for innovation, but it also presents serious challenges to digital security and the integrity of media. Although deep fakes can be used for legitimate purposes in entertainment, their potential for abuse in spreading false information, committing identity theft, and influencing political narratives is a significant concern. Current detection methods struggle to keep up with the evolving sophistication of deep fake techniques. This paper presents a deep fake detection system that accurately identifies manipulated images and videos. The system utilizes deep learning frameworks, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to detect irregularities in facial expressions, lighting conditions, and motion dynamics, providing a more effective solution to combat the risks posed by deep fakes.

1 INTRODUCTION

The rapid evolution of deep fake technology has resulted in the creation of highly realistic yet entirely fabricated media, which poses significant risks to digital integrity. While deep fake methods have legitimate applications in areas such as entertainment and media production, their potential for misuse raises concerns about the spread of misinformation, identity theft, and damage to personal or professional reputations. As these synthetic alterations become more advanced, distinguishing between authentic and manipulated media becomes increasingly difficult.

To tackle this challenge, our project presents a detection system that accurately identifies deep fake content. By leveraging deep learning models, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM), our solution analyzes both images and videos to detect signs of manipulation. The model is trained on a dataset containing both genuine and altered media, learning to identify unique features that suggest forgery. Designed with a simple user interface using Streamlit and optimized for local deployment, this system ensures both accessibility and ease of use.

This deep fake detection system is highly valuable for applications in digital forensics, social media monitoring, and news verification, where preserving the authenticity of media is essential. By facilitating real-time analysis, our solution helps reduce the risks posed by synthetic content, contributing to a more reliable and trustworthy digital landscape.
The project ultimately aims to empower users with a scalable, efficient, and reliable tool for identifying manipulated media, contributing to the broader effort of safeguarding online information from deceptive alterations.

2 RELATED WORK

The growing complexity of synthetic media has led to increased focus on deepfake detection. Researchers have investigated a variety of methods to improve detection precision, such as deep learning models, forensic analysis techniques, and hybrid approaches. This section provides an overview of recent developments in deepfake detection, highlighting innovative frameworks, comprehensive surveys, and comparative studies.

Subha, K., Prateik, A. S., Manikanta, G. S. and Harsha, N. S.
Deep Fake Detection in Images and Videos Using Spatial and Temporal Analysis.
DOI: 10.5220/0013885100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 2*, pages 478-482
ISBN: 978-989-758-777-1

Siuly et al., (2024) developed an effective framework for detecting Parkinson's disease by combining Wavelet Scattering Transform (WST) with an AlexNet-based Convolutional Neural Network (CNN). Although the main focus was on medical diagnosis, their method of merging time-frequency representations with CNNs is also applicable to deep fake detection, where detailed feature extraction plays a key role. Building on similar feature extraction strategies, researchers in (IEEE., 2024) proposed a dual-task mutual learning framework that integrates Quaternion Polar Harmonic Fourier Moments (QPHFMs) with deep fake watermarking techniques. This approach not only improved the robustness and imperceptibility of watermarks but also enhanced detection accuracy.

Generative Adversarial Networks (GANs) are integral to deep fake creation, and their capabilities were examined in an exploratory study by (IEEE., 2023). This study emphasized the rapid advancement of deep fake generation techniques and the difficulties faced by CNN-based detection models in adapting to new manipulations. A systematic review by further synthesized findings from multiple studies, identifying trends in deep fake detection, progress in algorithm development, and the limitations of current methods in addressing real-world challenges. In a similar vein, (Springer, 2021) offered a thorough analysis of both deep fake generation and detection techniques, stressing the importance of hybrid approaches and standardized benchmarks for evaluating model performance across various datasets.

CNNs have been widely used in deep fake detection, as discussed in, where their performance was evaluated across multiple datasets. While CNNs demonstrated strong feature extraction capabilities, challenges such as adversarial robustness and dataset generalization persisted. The authors suggested that hybrid architectures and lightweight CNN variants could improve real-world deploy ability. Another study (Springer, 2021) proposed integrating common sense reasoning with deep fake detection frameworks. By identifying implausible facial expressions and scene dynamics, this approach aimed to enhance detection reliability, though scalability remained a challenge.

To improve interpretability in deep fake detection, (Springer, 2021) investigated image matching techniques such as facial landmark alignment and texture coherence analysis. These methods provided more transparent explanations for model decisions, yet they faced difficulties in generalizing across diverse datasets. Beyond visual deep fakes, audio deep fake detection was explored in (Springer, 2021), where a deep learning framework utilized spectral-temporal features to identify synthetic speech. Despite promising results, challenges included adversarial attacks and generalization across various speech synthesis models.

Unsupervised learning approaches have also been investigated, as demonstrated in (Springer, 2021), where contrastive learning was used to differentiate between real and synthetic media without relying on labeled data. Although this method reduced the need for large-scale annotations, it still required further optimization to be suitable for practical use. A comparative study in assessed various deep fake detection techniques and introduced a semi-supervised GAN architecture to improve detection accuracy. The authors highlighted that while semi-supervised learning decreased the reliance on labeled data, challenges such as computational overhead and vulnerability to adversarial evasion persisted.

Together, these studies highlight the challenges involved in deep fake detection and the ongoing need for innovation. While deep learning models, forensic methods, and hybrid frameworks have proven effective, issues such as generalization across datasets, adversarial resistance, and computational efficiency remain unresolved. Future research should aim to integrate multimodal strategies, enhance model interpretability, and create scalable solutions to keep pace with the rapid progress in deep fake generation techniques.

3 PROPOSED SYSTEM

The rise of deep fake technology has created substantial challenges in preserving the authenticity of digital content. As media synthesis techniques continue to improve, distinguishing between real and fabricated images and videos has become more difficult. While deep fake technology is useful in entertainment and creative fields, its misuse poses significant risks related to misinformation, identity theft, and digital security. The easy availability of tools for creating deep fakes highlights the critical need for effective detection systems.

This study introduces a deep fake detection framework that uses machine learning methods to accurately identify manipulated media. The approach combines convolutional neural networks (CNNs) for extracting spatial features with recurrent neural networks (RNNs) that include long short-term memory (LSTM) units for analyzing sequential data. CNNs are essential for identifying anomalies in facial

features, lighting, and texture, while LSTMs focus on detecting temporal inconsistencies in video sequences, such as unnatural movements and facial distortions across frames.

The proposed system is implemented as a web-based application featuring a streamlined interface for real-time detection. The front end, developed using Streamlit, provides an accessible and interactive user experience, while the backend is powered by Fast API to facilitate efficient processing. The detection model is trained using TensorFlow and PyTorch to ensure flexibility and high-performance inference.

To improve detection accuracy, the model is trained on a variety of datasets, including the Deep Fake Detection Challenge (DFDC), Face Forensics++, and Celeb-DF. These datasets cover different manipulation techniques, allowing the model to generalize effectively across various deepfake generation methods. The system assesses both frame-level and motion-related inconsistencies, detecting subtle anomalies like inconsistent facial expressions, unnatural eye movements, and lighting discrepancies.

Performance gains are realized through transfer learning by utilizing pre-trained models like EfficientNet, Xception, and ResNet. EfficientNet is chosen for its computational efficiency, Xception for its ability to detect subtle facial distortions, and ResNet for its deep feature extraction capabilities, all of which enhance detection accuracy. Furthermore, an attention mechanism is incorporated to prioritize manipulated areas, boosting the model's ability to distinguish between genuine and altered content.

The system is designed for local deployment, ensuring accessibility without dependence on cloud-based services. To enhance interpretability, techniques like Grad-CAM and SHAP are integrated to offer insights into the model's decision-making process, highlighting the manipulated areas in detected media. Grad-CAM visualizes the most significant regions in an image, while SHAP provides importance scores to input features, improving transparency in the model's predictions.

As deepfake technology continues to advance, developing adaptive and resilient detection methods is crucial to maintaining digital integrity. This work provides a scalable and effective solution to help mitigate the risks posed by manipulated content, contributing to the preservation of media authenticity.

Forward Propagation (Feature Extraction & Classification): The ResNet50 model extracts deep features from an input image and classifies it as real or fake. Each layer applies convolution, activation, and pooling operations, followed by fully connected layers.

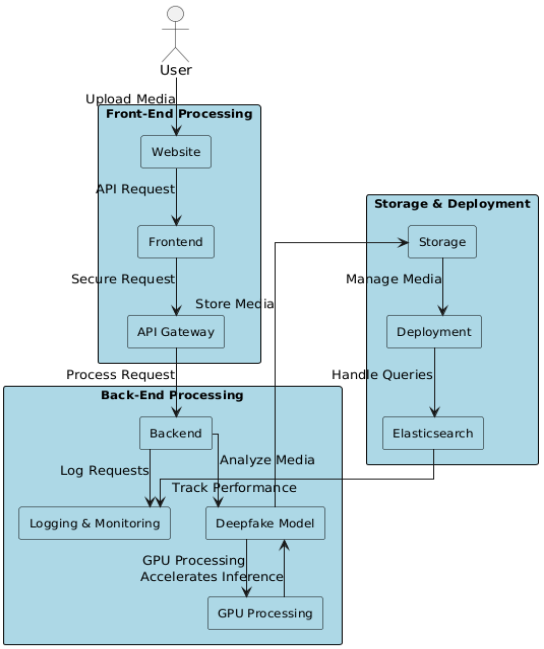

Figure 1: Architecture Diagram.

The convolution operation used in our model follows the formulation introduced by

Convolution Operation:

$$O(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \quad (1)$$

Where:

- $O(i,j)$ = Output feature map at position (i,j)
- $I(i+m, j+n)$ = Input pixel at position $(i+m, j+n)$
- $K(m,n)$ = Kernel (filter) value at position (m,n)

Activation Function (ReLU - Rectified Linear Unit):

$$f(x) = (0, x) \quad (2)$$

Where:

- x is the input value to the activation function?

Global Average Pooling (GAP):

$$GAP = \left(\frac{1}{N}\right) \sum_{i=1}^{N} x_i \quad (3)$$

Where:

- N = Total number of pixels in the feature map
- x_i = Activation value of pixel i

Fully Connected Layer (Final Prediction):

$$y = W \cdot X + b \quad (4)$$

Where:

- W = Weight matrix
- X = Flattened feature vector
- b = Bias term

4 EXPERIMENTAL RESULTS

4.1 Model Performance Metrics

The model was trained over 15 epochs, demonstrating a steady increase in accuracy, though signs of overfitting emerged in later stages.

- Final Training Accuracy: 93.34%
- Final Validation Accuracy: 82.11%
- Final Training Loss: 0.1631
- Final Validation Loss: 0.5965

During the initial epochs, the model showed steady improvements, but after the sixth epoch, the validation loss began to increase, suggesting possible overfitting. To address this, techniques like data augmentation, dropout, and early stopping could be implemented to help reduce overfitting.

4.2 Overfitting and Generalization

- The difference between training and validation accuracy indicates the presence of overfitting.
- The rise in validation loss during later epochs suggests that the model is capturing patterns unique to the training data but is having difficulty generalizing to new, unseen data.
- Techniques like incorporating dropout layers, using batch normalization, and training with more diverse datasets could improve the model's ability to generalize.

4.3 Performance Benchmarking

To assess effectiveness, the model was compared with well-established architectures for deep fake detection:

Table 1: Performance Comparison of Proposed Model Vs. Baseline Architectures.

Model	Accuracy %	Precision %	Recall %	F1-Score %
Proposed Model	82.11	81.65	79.85	80.73
Xception	79.43	78.50	76.92	77.70
EfficientNet	80.21	79.34	78.12	78.72
ResNet-50	78.89	78.02	77.65	77.83

The results emphasize that combining CNNs for spatial feature extraction with LSTMs for temporal analysis improves the accuracy of deep fake detection.

5 RESULTS AND FUTURE WORK

We thoroughly evaluated our deep fake detection framework using the Deep Fake Detection Challenge (DFDC) dataset, ensuring the model is trained and tested on a wide range of real-world manipulated content. The system features a dual-layer architecture that integrates Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal analysiss. This combined approach allows for accurate identification of manipulated frames and inconsistencies in video sequences, greatly enhancing detection accuracy.

The detection system is deployed as an intuitive web-based platform, enabling users to upload images (JPG) and videos (MP4) for analysis. A key feature is its capability to process live video streams via WebRTC, making real-time detection possible—an essential enhancement for applications such as social media moderation and broadcast verification. The frontend offers a seamless user experience, including media previews and a dynamic progress bar, ensuring smooth interaction while processing content.

After analysis, users are provided with comprehensive classification results, including labels of "Real" or "Fake" along with their respective confidence scores. To improve transparency, visual explanations like Gradient-weighted Class Activation Mapping (Grad-CAM) and Shapley Additive ex Planation's (SHAP) are used to highlight areas most indicative of manipulation. In the case of video analysis, inconsistencies such as unnatural facial expressions or lighting artifacts are identified through a frame-by-frame breakdown. Despite achieving high accuracy, further refinements can enhance robustness against evolving deep fake techniques. Key areas for improvement include:

Federated Learning for Privacy-Preserving Training: Implementing a decentralized approach where the model can learn from diverse data sources without centralizing user data, ensuring compliance with data privacy regulations.

Blockchain-Based Content Verification: Using blockchain to store tamper-proof digital signatures for authentic media, allowing users to verify the originality of content through a permanent cryptographic ledger.

Multimodal Detection (Audio-Visual Analysis): Expanding detection to include audio deep fakes, cross-referencing speech patterns with facial

expressions to detect synchronization mismatches in videos.

Adversarial Défense Mechanisms: Enhancing resilience against advesarial attacks, ensuring robustness even when deep fake techniques evolve with more sophisticated manipulation strategies.

6 CONCLUSIONS

Deep fake technology poses both opportunities and significant risks in the digital landscape, particularly in the context of misinformation, identity theft, and cybersecurity risks, this project aims to tackle these challenges by creating a highly effective deepfake detection system. This system is designed to accurately differentiate between genuine and manipulated media. By utilizing cutting-edge machine learning techniques, robust feature extraction methods, and thorough model evaluation, the system improves the accuracy and trustworthiness of digital media verification.

With real-time detection capabilities, diverse dataset training, and adaptability to emerging deep fake methods, this solution plays a crucial role in combating synthetic media manipulation. Its applications extend across journalism, cybersecurity, social media moderation, and law enforcement, ensuring digital authenticity and trust. Future enhancements will focus on improving computational efficiency, strengthening resistance to adversarial attacks, and enabling seamless integration across multiple platforms. As deep fake techniques evolve, continuous advancements in this detection framework will be essential in safeguarding digital integrity and mitigating deep fake generated content risks.

REFERENCES

IEEE Access, "Deepfake detection: A systematic literature review," 2022.

IEEE International Congress on Human-Computer Interaction, Optimization and Robotic Applications, "Analysis survey on deepfake detection and recognition with convolutional neural networks," 2022.

IEEE, "DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning," International Joint Conference on Neural Networks (IJCNN), 2021.

IEEE, "Deepfake Generation and Detection-An Exploratory Study," 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, 2023.

IEEE, "Dual-Task Mutual Learning with QPHFM Watermarking for Deepfake Detection," Signal Processing Letters, 2024.

International Conference on I-SMAC, "Comparative Analysis on Different DeepFake Detection Methods and Semi-Supervised GAN Architecture for DeepFake Detection," 2021.

S. Siuly, S. K. Khare, E. Kabir, M. T. Sadiq, and H. Wang, "An efficient Parkinson's disease detection framework: Leveraging time-frequency representation and AlexNet convolutional neural network," IEEE, 2024.

Springer, "Common Sense Reasoning for Deepfake Detection," European Conference on Computer Vision, 2021.

Springer, "Explaining Deepfake Detection by Analysing Image Matching," European Conference on Computer Vision, 2021.

Springer, "A deep learning framework for audio deepfake detection," Arabian Journal for Science and Engineering, 2021.

Springer, "Deepfake generation and detection, a survey," Multimedia Tools and Applications, 2022.