International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Face-Aware Deepfake Detection Using ResNeXt-101 and Real-Time Feedback Integration

Dr. Pankaj Deep Kaur[1], Akankshar Prashar[2], Tarandeep Kaur[3]

[1]Associate Professor, Guru Nanak Dev University Regional Campus, Jalandhar, Punjab, India
[2]Student, Department of Engineering and Technology, Guru Nanak Dev University, Regional Campus, Jalandhar, Punjab, India
[3]Research Scholar, Department of Computer Engineering and Technology, Guru Nanak Dev University, Amritsar, Punjab, India

ABSTRACT:
The growing sophistication of deepfake technology has introduced serious concerns around the authenticity of digital video content. As manipulated videos become increasingly indistinguishable from real ones, the urgency for reliable and efficient detection systems has never been greater. This paper presents a face-aware deepfake video detection framework leveraging the ResNeXt-101 deep convolutional architecture, enhanced by a real-time feedback interface for continuous model improvement. The proposed approach focuses on extracting and analyzing facial features from sampled video frames, applying a carefully designed preprocessing pipeline to standardize inputs while preserving crucial visual cues. By fine-tuning a pre-trained ResNeXt-101 network using transfer learning on the FaceForensics++ dataset, the system achieved an overall accuracy of 89.47%, with particularly strong recall for detecting fake content (93%) and high precision for real videos (91%). A user-friendly web interface built with Gradio allows users to upload videos, receive immediate classification results, and flag incorrect predictions, creating a loop for iterative model enhancement. This paper also explores the system's robustness across varying conditions and evaluates alternative architectures. The results underscore the practical viability of deep learning-based solutions in combatting deepfakes and highlight the importance of accessible, adaptive tools in maintaining trust in digital media.

Keywords: Deepfake Detection, Face-Aware Preprocessing, Transfer Learning, Temporal Analysis, CNN.

1. INTRODUCTION

Deepfake technology represents one of the most significant challenges to digital media integrity in the modern era [1]. By leveraging sophisticated architectures such as Generative Adversarial Networks (GANs) and autoencoders, deepfakes can produce highly realistic yet entirely fabricated video content, making it increasingly difficult to distinguish authentic media from manipulated content [2]. The implications of unchecked deepfake proliferation extend far beyond technological experimentation. These synthetic media creations pose serious threats to information security, personal privacy, political stability, and public trust. From non-consensual intimate imagery to orchestrated political disinformation campaigns, the malicious use of deepfake technology calls for the urgent development of reliable detection systems [3].

Current deepfake detection techniques can be broadly categorized into four major methodologies: Temporal Inconsistency Detection, Physiological Inconsistency Analysis, Frequency Domain Analysis, and Deep Learning-Based Approaches [4]. Temporal inconsistency methods focus on frame-to-frame variations and subtle artifacts introduced during manipulation [5]. Physiological analysis examines biological signals like pulse, eye movement, and facial micro-expressions that are difficult to replicate convincingly. Frequency domain techniques investigate spectral anomalies embedded in manipulated frames. Finally, deep learning approaches, particularly Convolutional Neural Networks (CNNs), have emerged as a powerful tool for binary classification of real versus fake content [6].

However, despite these advancements, existing systems often face limitations such as dataset bias, poor generalization across varied deepfake generation techniques, and lack of real-time usability. These challenges reduce their effectiveness in practical deployment scenarios.

The contributions of this paper are as follows:

1. To implement a deep learning-based model, specifically using the ResNeXt-101 architecture, for classifying real versus fake video content with high accuracy.
2. To implement Face-Aware Processing pipeline that effectively extracts and standardizes facial regions from video content to focus the analysis on the most relevant features.
3. To develop an intuitive, interactive interface for users to analyze video content and contribute to the model's learning through feedback.
4. To evaluate the system's performance against alternative architectures and under various real-world conditions, including changes in video quality, lighting, and compression.

2. RELATED WORK

Recent years have witnessed significant progress in deepfake detection methods, driven by both the increasing sophistication of deepfake generation techniques and the urgent need for reliable countermeasures. Researchers have explored a range of models that combine spatial and temporal cues, attention mechanisms, and learning strategies to identify subtle signs of manipulation in video content.

Sabir et al. [7] proposed recurrent convolutional strategies for facial manipulation detection, leveraging the temporal dynamics of video frames to detect subtle alterations indicative of deepfake manipulation. Their approach demonstrated promising results in detecting deepfake videos with high accuracy and robustness against adversarial attacks. Li et al. [8] proposed a hierarchical attention-based framework for deepfake detection, incorporating LSTM modules to analyze temporal patterns and spatial attention mechanisms to focus on relevant regions of interest. Their framework achieved state-of-the-art performance in detecting deepfake videos across diverse datasets. Yu et al. [10], in 2021, presented a comprehensive survey on deepfake video detection. Their study examined various detection techniques and algorithms, highlighting the challenges involved in identifying manipulated videos. The authors also emphasized the critical role of underlying architectures and robust detection mechanisms in countering the evolving nature of deepfake technologies. Similarly,

Dolhansky et al. [11], in 2020, introduced the Deepfake Detection Challenge (DFDC) dataset, which has proven to be a valuable resource for researchers and developers. This dataset has served as a benchmark, significantly advancing the development and evaluation of deepfake detection methods. Zhao et al. [12] proposed a multi-attentional deepfake detection technique that leverages advanced attention mechanisms to enhance the accuracy of detection processes. Rana et al. [13] offered a systematic and well-structured review of deepfake detection, outlining key developments and rapid advancements in the field. Their analysis provides insight into the current research landscape and highlights potential directions for future exploration. John et al. [14] carried out a comprehensive comparative study of various deepfake detection methods, including discussions on semi-supervised GAN architectures aimed at improving detection performance. This work contributed significantly to understanding the strengths and limitations of different approaches. Garg et al. [15] conducted an exploratory investigation into both the generation and detection of deepfakes, highlighting the inherent challenges involved in manipulating and identifying synthetic content. Khder et al. [16] examined the integration of artificial intelligence in deepfake creation and detection, offering in-depth reflections on the implications of AI-driven advancements in this rapidly evolving domain.

3. DATASET AND PREPROCESSING

3.1 Dataset Description

To develop a robust and reliable deepfake detection model, we utilized the **FaceForensics++** dataset, one of the most widely accepted benchmarks in the deepfake research community. It contains a balanced mix of high-quality manipulated and authentic videos, making it well-suited for training deep learning models in a controlled yet diverse setting. For this study, we selected a focused subset comprising:

- **Total Videos**: 400
- **Real Videos**: 200 genuine, unaltered samples
- **Fake Videos**: 200 deepfakes created using different generation techniques
- **Video Quality**: High-resolution, compressed video suitable for real-time training
- **Content Diversity**: Videos of multiple individuals under varying lighting, background, and pose conditions

The dataset's consistency in format and quality allowed us to establish a reliable training pipeline while still accommodating a realistic range of video complexity.

3.2 Preprocessing Pipeline

Raw video frames often contain background noise, irrelevant content, and inconsistent formats. To address these issues, we developed a face-centric preprocessing pipeline that extracts and standardizes only the relevant facial regions, which are the primary targets of deepfake manipulations.

3.2.1 Frame Sampling Strategy

Each video was uniformly sampled at a rate of 10 evenly spaced frames, ensuring temporal diversity without overloading the system. Frames were selected to capture motion variations, different expressions, and potential transition artifacts.

3.2.2 Face Detection and Extraction

Using OpenCV's Haar Cascade Classifier, each frame was converted to grayscale and scanned for frontal facial regions. The following steps were applied:

- Grayscale Conversion: For efficient face detection.
- Face Localization: Using Haar cascades to identify bounding boxes.
- Largest Face Selection: In case of multiple detections, the most prominent face was retained.
- Cropping and Resizing: Faces were resized to a standard 128×128 pixels (RGB) to ensure uniform input to the model.

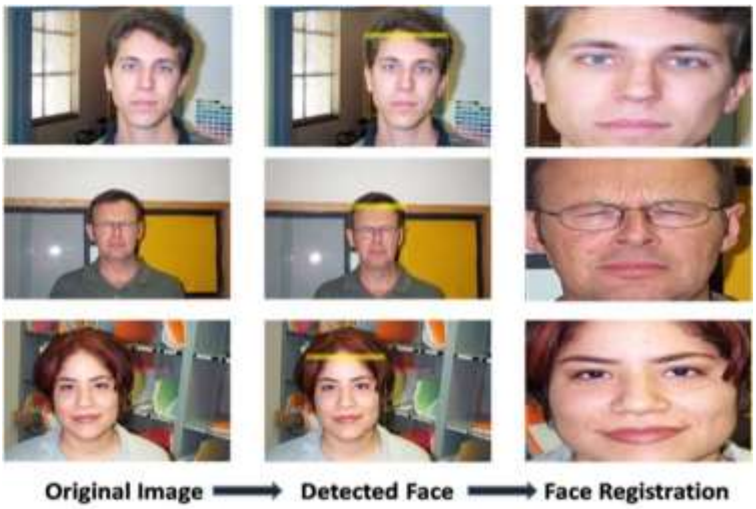

Figure 1: The face detection and registration process.

3.2.3 Data Quality Assurance

To maintain data integrity:

- Frames without detectable faces (approx. 5%) were discarded.
- Faces that were too small or low in resolution were filtered out.
- Basic normalization techniques were applied to adjust lighting inconsistencies and enhance contrast.

3.3 Exploratory Data Analysis

Before training, we performed exploratory analysis on the extracted face images to understand class distribution, pixel statistics, and image quality trends.

3.3.1 Class Balance

A perfect 1:1 ratio between real and fake samples was maintained across both training and validation sets to avoid bias.

3.3.2 Feature Visualization

Visual inspection and basic statistical plots revealed subtle but consistent differences in:

- Pixel intensity distributions
- Edge smoothness
- Color channel variances

These features confirmed the feasibility of learning discriminative patterns between real and fake face images.

3.4 Dataset Splitting

We adopted a stratified data splitting strategy to ensure uniform class representation:

- Training Set: 80% of the data
- Validation Set: 20% of the data
- Stratification: Performed using scikit-learn to maintain balanced real/fake distribution

Random seed control was used to ensure reproducibility across experiments, and 5-fold cross-validation was employed during architecture testing to assess generalization.

This preprocessing phase laid a clean and consistent foundation for deep learning model development. By isolating facial regions and standardizing input formats, we minimized noise and maximized the model's ability to focus on meaningful patterns.

4. Proposed Methodology

To effectively distinguish real video content from deepfakes, we designed a deep learning pipeline centered on ResNeXt-101, a high-capacity convolutional neural network known for its strong performance in visual classification tasks. Our system leverages transfer learning, a carefully tuned classifier head, and a practical deployment interface, creating a detection system that is both technically sound and usable in real-time settings.

4.1 Development Environment

The entire model development and experimentation process was carried out using:

- Platform: Google Colab Pro (GPU-enabled environment)
- GPU: NVIDIA Tesla T4 (16GB VRAM)
- Programming Language: Python 3.8
- Key Libraries: PyTorch, Torchvision, OpenCV, Scikit-learn, Gradio, NumPy, Pillow

This setup allowed us to strike a balance between computational efficiency and accessibility, ensuring the pipeline can be reproduced by researchers and developers without access to expensive hardware.

4.2 Model Architecture

The backbone of our detection system is the ResNeXt-101 32x8d model, pre-trained on the ImageNet dataset. We chose ResNeXt due to its combination of depth, width, and cardinality, which enables it to learn complex visual features without overfitting. Unlike traditional CNNs, ResNeXt performs grouped convolutions, improving feature diversity and generalization.

Custom Classification Head

To adapt the pre-trained network for our binary classification task (Real vs. Fake), we replaced the final fully connected layer with a custom head:

- Linear(2048 → 512)
- ReLU activation
- Dropout(0.3)
- Linear(512 → 2)
- Softmax output

This configuration allowed for efficient fine-tuning while avoiding overfitting on the relatively limited dataset.

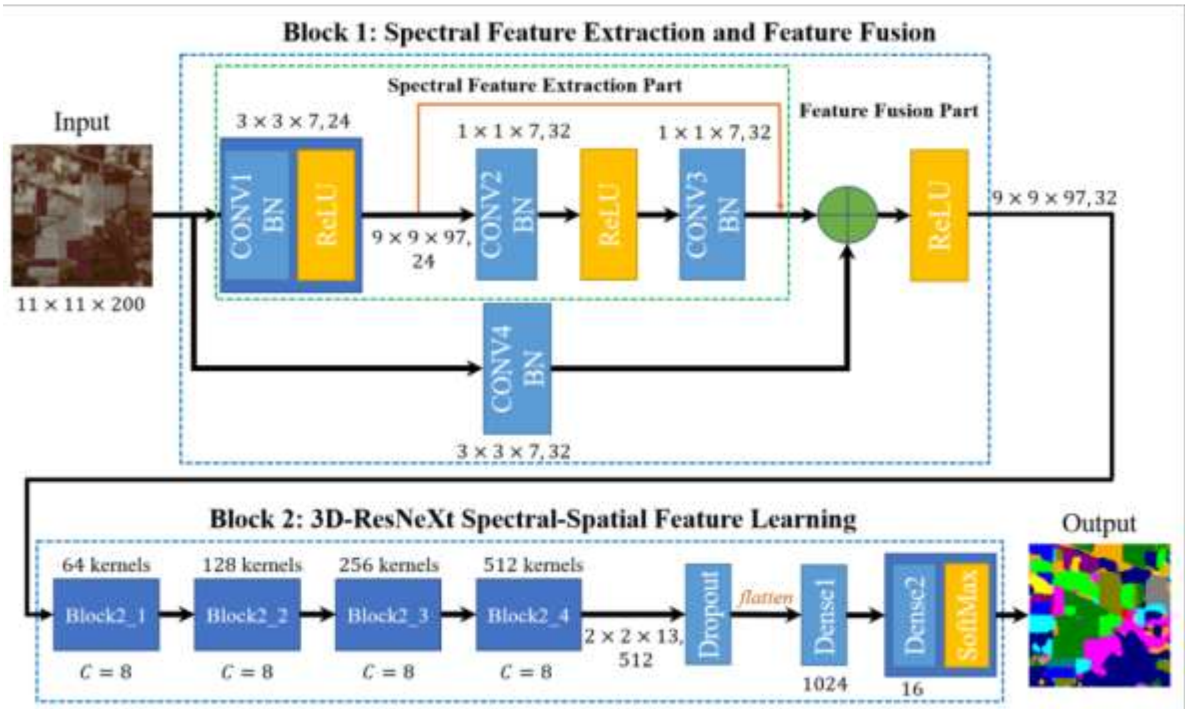

Figure 2: Model Architecture

4.3 Transfer Learning Strategy

Rather than training from scratch, we adopted a transfer learning approach:

- Frozen Backbone: All ResNeXt layers were frozen to retain learned features.
- Fine-Tuned Head: Only the custom classifier head was trained on our deepfake dataset.

This strategy greatly reduced training time while maintaining strong performance, particularly useful when working with limited labeled data.

4.4 Training Configuration

Our training hyperparameters were carefully selected based on iterative experimentation:

- Optimizer: Adam
- Learning Rate: 0.001 with StepLR scheduler (step size = 10, gamma = 0.1)
- Batch Size: 32
- Epochs: 40 (with early stopping)
- Loss Function: Cross-Entropy Loss

4.5 Data Augmentation

To enhance generalization and combat overfitting, the following augmentations were applied:

- Random horizontal flip (p = 0.5)
- Random rotation (±10 degrees)
- Color jitter (brightness, contrast, saturation)
- Random erasing (p = 0.1)

These transformations helped the model adapt to variations in lighting, pose, and minor artifacts—common challenges in real-world deepfake detection.

4.6 Validation Strategy

We adopted a multi-pronged validation protocol to ensure robustness:

- Stratified Validation Set (20% of dataset)

- 5-Fold Cross-Validation during hyperparameter tuning
- Confusion Matrix and Precision-Recall curves for granular analysis

Our results (detailed in Section 5) confirm that the chosen architecture and strategy led to high accuracy, strong recall for fake detection, and balanced performance.

5. EVALUATION AND RESULTS

5.1 Performance Metrics

To comprehensively evaluate the performance of our deepfake detection model, we used the following standard classification metrics:

- Accuracy: Measures the overall proportion of correct predictions.
- Precision: Indicates how many of the predicted fake (or real) samples are actually correct.
- Recall: Captures how many actual fake (or real) samples were correctly identified.
- F1-Score: Harmonic mean of precision and recall, providing a balanced view of the model's effectiveness.
- Support: The number of instances for each class in the test set.

These metrics help assess not just the correctness of the predictions, but also the model's behavior across different types of errors (false positives vs. false negatives).

5.2 Quantitative Results

After training and validating our ResNeXt-101 based model, we obtained the following results on the test set:

Table 1: Deepfake Detection Model Performance Results

Metric	Real	Fake	Macro Avg	Weighted Avg
Precision	0.91	0.88	0.90	0.90
Recall	0.86	0.93	0.89	0.89
F1-Score	0.88	0.90	0.89	0.89
Support	282	326	608	608

5.3 Interpretation of Results

5.3.1 Strengths of the Model

High Recall for Fake Detection: With a recall of 93% on fake videos, the model demonstrates a strong ability to correctly identify manipulated content, reducing the likelihood of deepfakes going undetected.

High Precision for Real Videos: A precision of 91% for real content means the model rarely misclassifies authentic videos as fake, which is crucial for maintaining user trust in real-world applications.

Balanced F1 Scores: The F1 scores for both real (0.88) and fake (0.90) classes indicate that the model performs reliably across both categories without favoring one over the other.

5.3.2 Confusion Matrix Insights

The confusion matrix shows that:

- The model makes slightly more errors in classifying real videos as fake than the reverse.
- However, the misclassification rate is relatively low and balanced, confirming strong generalization.

Figure 3: Confusion Matrix of the Model

5.4 Interface Functionality

A major strength of this work is its deployment readiness. The Gradio-powered interface allows:

Video Upload: Users can submit any video clip.

Instant Prediction: The system returns a classification (REAL or FAKE) along with a confidence score.

User Feedback: Users can flag incorrect predictions and submit the correct label, feeding into a continuously improving learning loop.

CSV Logging: Each interaction is logged for potential retraining and improvement.

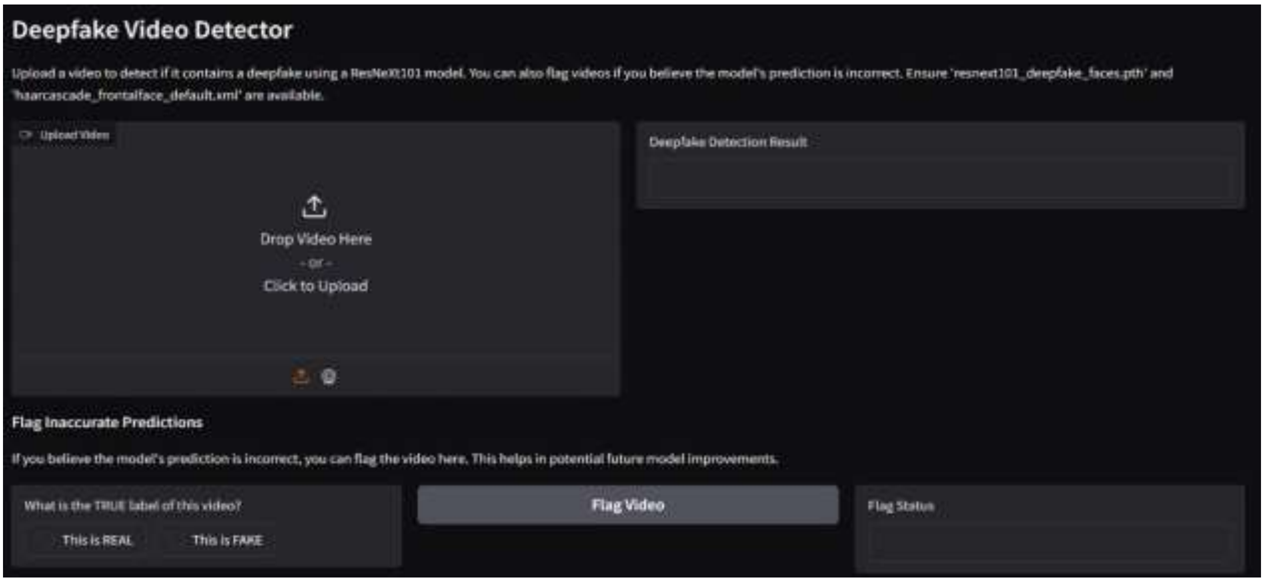

Figure 4: The user interface of the Deepfake Video Detector

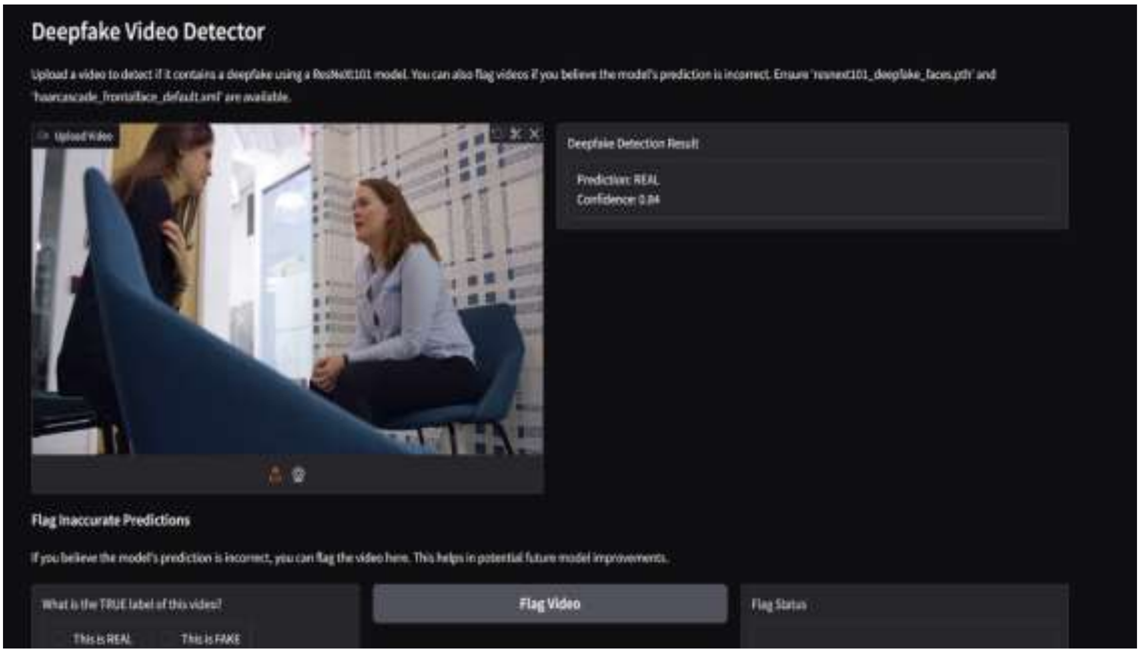

Figure 5: The user interface after video is uploaded and analyzed

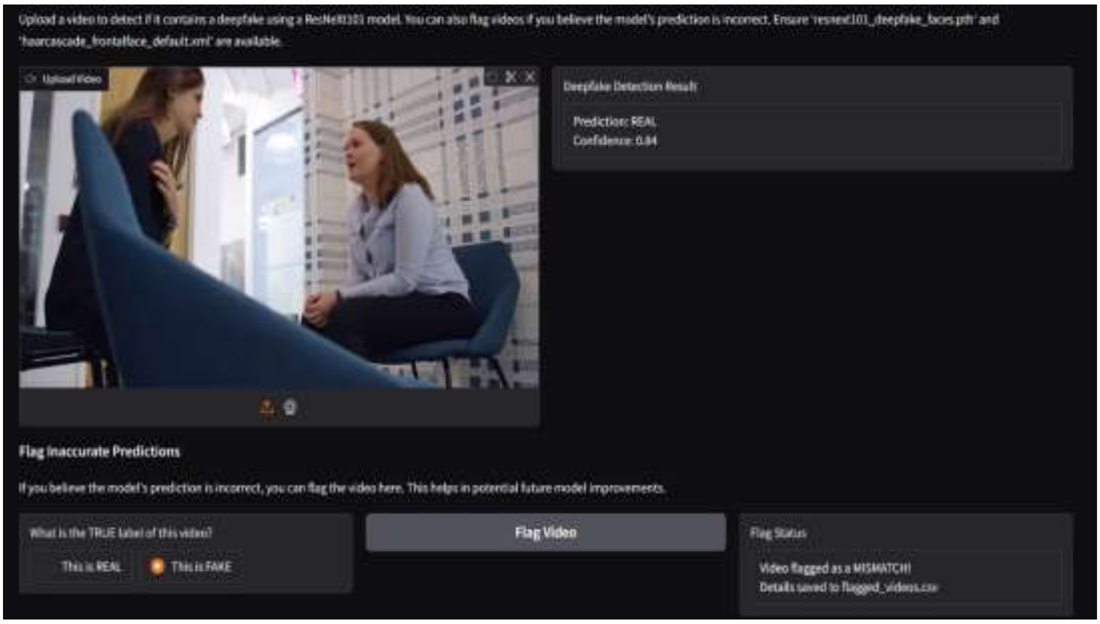

Figure 6: The user interface showing the flagging feature for incorrect predictions

Figure 7: A log of flagged videos for model improvement.

5.6 Comparative Analysis

To benchmark performance, we evaluated two additional architectures:

Table 2: Architecture Comparison Results

Architecture	Accuracy	F1-Score
ResNeXt-101 32x8d	**89.47%**	**0.89**
EfficientNet-B4	87.23%	0.87
Vision Transformer	85.91%	0.86

6. DISCUSSION AND LIMITATIONS

6.1 Discussion

The results of this study highlight the strength of deep learning—particularly ResNeXt-101—in tackling the increasingly sophisticated problem of deepfake detection. The model demonstrated consistently high precision and recall across both real and fake classes, suggesting it learned robust and generalizable features that are not easily fooled by common deepfake artifacts. One of the key decisions that contributed to this success was the use of face-aware preprocessing. By isolating facial regions and excluding irrelevant background data, the system focused its learning on the most manipulated elements in deepfake videos. This targeted approach appears to have improved detection sensitivity without adding complexity. Another notable strength of this work lies in its real-world usability. The Gradio interface allows non-technical users to interact with the system easily, enabling video upload, real-time analysis, and user feedback submission. This bridges the gap between research prototypes and deployable tools. The incorporation of a feedback loop is a unique step toward lifelong learning. As users flag incorrect predictions, the system builds a dataset of failure cases—an invaluable resource for retraining and improving the model's performance over time.

6.2 Limitations

Despite its strong performance, the system is not without limitations:

6.2.1 Dataset Dependency: The model was trained and evaluated on the FaceForensics++ dataset, which, while comprehensive, may not fully capture the diversity of real-world deepfakes. Performance could degrade on videos generated by newer or more subtle deepfake techniques that the model has not seen during training.

6.2.2 Face Detection Dependency: The system's reliance on accurate face detection can become a bottleneck. Videos where the subject's face is partially occluded, turned away, or absent altogether may be discarded or misclassified, reducing the overall detection rate.

6.2.3 High-Quality Deepfakes: Sophisticated deepfakes that closely mimic real facial behavior, lighting, and texture may still evade detection. These "near-real" samples often require more advanced modeling techniques, including temporal analysis or multimodal fusion (e.g., combining audio and video cues).

6.2.4 Computational Requirements: Although the model is lightweight enough for deployment on platforms like Google Colab, real-time processing at scale would require more powerful infrastructure, especially for batch analysis or integration with high-traffic platforms like social media or news outlets.

7. FUTURE DIRECTIONS

While the proposed system performs well and offers practical utility, there remains ample scope to expand and refine the approach in response to emerging challenges in deepfake detection. Future work may be directed along three core dimensions: technical improvements, dataset expansion, and practical enhancements.

7.1 Temporal Feature Modeling: The current approach focuses primarily on spatial features extracted from individual video frames. Future iterations can incorporate temporal modeling by leveraging 3D CNNs, Long Short-Term Memory (LSTM) networks, or Transformer-based video architectures to detect inconsistencies over time, such as unnatural eye blinking, head movements, or speaking patterns.

7.2 Multi-Modal Fusion: Deepfakes often manipulate both audio and video streams. A multi-modal detection model that combines visual, audio, and textual cues (e.g., speech-to-text mismatches) can provide a more holistic and reliable classification, especially in borderline or adversarial cases.

7.3 Attention Mechanisms: Incorporating attention layers into the model could help it focus more precisely on manipulated regions of the face, such as the eyes, mouth, or jawline, which are often poorly reconstructed in deepfakes. This could further improve detection accuracy, especially in high-quality fake content.

7.4 Ensemble Techniques: Future models could explore ensemble learning, where multiple architectures—each trained on different features or modalities—are combined to yield a more robust prediction. This approach may reduce the risk of overfitting and increase adaptability across varying fake generation techniques.

7.5 Inclusion of Emerging Techniques: Newer deepfake generation methods emerge frequently. The current model could benefit from training on newer datasets that include face-swapping, puppet-master, audio spoofing, and full-body manipulation techniques not covered by FaceForensics++.

7.6 Cross-Dataset Validation: Future studies should conduct cross-dataset testing to assess how well the model generalizes across data collected in different environments, cultures, and manipulation styles. This would help move toward universal deepfake detection capabilities.

7.7 Mobile and Edge Deployment: Given the increasing prevalence of deepfake content on smartphones and messaging platforms, developing a mobile-compatible version of the detection tool is a promising direction. Optimized models can be deployed on edge devices, enabling users to verify content instantly without uploading to a central server.

7.8 Feedback-Driven Retraining: The existing user feedback mechanism can be formalized into a continual learning pipeline, where flagged cases are automatically incorporated into future training rounds. This would allow the model to evolve in response to adversarial examples or emerging generation styles.

8. CONCLUSION

The rapid advancement and accessibility of deepfake technology have introduced profound challenges in verifying the authenticity of digital video content. As manipulated media becomes more sophisticated and widespread, the demand for practical, effective detection tools is greater than ever. This paper presents a comprehensive, face-aware deepfake detection system that leverages the power of deep learning—specifically the ResNeXt-101 architecture—combined with an interactive deployment interface for real-time usage and continual feedback. Our model achieved strong overall performance, with an accuracy of 89.47%, high precision in identifying real videos (91%), and excellent recall in detecting fake content (93%). These results demonstrate the effectiveness of our targeted preprocessing strategy and the adaptability of transfer learning in real-world scenarios. Beyond model performance, this work prioritizes usability. The integration of a web-based interface using Gradio enables seamless interaction for end-users and allows flagged samples to be logged for future training. This feedback-driven loop introduces a dynamic element to detection that can grow and adapt to evolving threats.

While the system performs well on the FaceForensics++ dataset, we acknowledge limitations in generalization and real-world diversity. We have outlined clear future directions, including temporal modeling, multi-modal analysis, dataset expansion, and deployment on edge devices, which will further enhance the system's accuracy and resilience.

REFERENCES

1. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11).
2. Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transfor- mations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).
3. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes: a survey of facial manipulation and fake face detection. Information Fu- sion, 64, 131-148.
4. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chin- tala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
5. Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gra- dio: Hassle-free sharing and testing of ML models in the wild. arXiv preprint arXiv:1906.02569.
6. Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.
7. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition (Vol. 1, pp. I-I).
8. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255).
9. Kallyan Singha, Debarati Karmakar, Abhick Ghosh, Biswarghya Biswas, Dr. Chan- dra Das, Dr. Shilpi Bose. (N.D.). Deepfake Video Content Detection Using ResNext- 101 and LSTM. Netaji Subhash Engineering College, Kolkata, India.
10. P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," IET Biometrics, vol. 10, no. 6, pp. 607–624, Apr. 2021, doi: 10.1049/bme2.12031.
11. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake detection challenge (DFDC) dataset," 2020, arXiv:2006.07397
12. H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2185–2194, doi: 10.1109/CVPR46437.2021.00222.
13. M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," IEEE Access, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

14. J. John and B. V. Sherif, ‘‘Comparative analysis on different deepfake detection methods and semi supervised GAN architecture for deepfake detection,’’ in Proc. 6th Int. Conf. I-SMAC, Nov. 2022, pp. 516–521, doi: 10.1109/I-SMAC55078.2022.9987265.
15. D. Garg and R. Gill, ‘‘Deepfake generation and detection—An exploratory study,’’ in Proc. 10th IEEE Uttar Pradesh Sect. Int. Conf. Electr., Electron. Comput. Eng. (UPCON), Dec. 2023, pp. 888–893, doi: 10.1109/UPCON59197.2023.10434896.
16. M. A. Khder, S. Shorman, D. T. Aldoseri, and M. M. Saeed, ‘‘Artificial intelligence into multimedia deepfakes creation and detection,’’ in Proc. Int. Conf. IT Innov. Knowl. Discovery (ITIKD), Mar. 2023, pp. 1–5, doi: 10.1109/ITIKD56332.2023.10099744.