Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 6
June 2025

Improved Hybrid Convolutional Neural Network Approach Based on Deep Learning for Deepfake Detection

Tejal Nichit†, Alekhya Muthineni†, Shaily Patel†, Siddhi Borhade†, Dr.Priyank Jain†

* Computer Science and Engineering,

Indian Institute of Information Technology Pune, 411041, Maharashtra, India.

†These authors contributed equally to this work.

Abstract

The rapid advancement of deep learning has enabled the creation of highly realistic deepfake media, raising concerns in cybersecurity, misinformation, and digital forensics. Traditional deepfake detection approaches primarily rely on convolutional neural networks (CNNs), which are often vulnerable to adversarial attacks and struggle to generalize across datasets. This study proposes a hybrid deepfake detection framework integrating ResNet50, MesoNet4, and an Eye Movement Analysis module to improve detection accuracy and robustness. ResNet50 is chosen for its ability to extract high-level spatial features, while MesoNet4 is optimized for detecting low-resolution manipulations. Additionally, physiological cues such as blink rate and gaze shift inconsistencies are leveraged to enhance detection effectiveness. Experimental results on multiple datasets (Deepfake-and-Real Images, UADFV, FaceForensics++) demonstrate that our model achieves 97.61% accuracy, out performing standalone CNN's. Grad-CAM visualization further enhances model interpretability, making the detection process more transparent. This research highlights the importance of combining CNN-based feature extraction with physiological-based detection for more robust and explainable deepfake forensics.

Keywords: Deepfake Detection, Convolutional Neural Network (CNN), Deep Learning, Hybrid Model, Fake Media Identification

1. Introduction

The rapid advancement of deep learning has enabled the generation of highly realistic deepfake images and videos, raising significant concerns about misinformation, identity fraud, and cybersecurity threats. Conventional deepfake detection methods primarily rely on pixel-based inconsistencies but often struggle with adversarial robustness and generalization across diverse datasets.

This study proposes a hybrid deepfake detection framework integrating ResNet50, MesoNet4, and an Eye Movement Analysis module. ResNet50 is utilized for extracting high-level spatial features from high-resolution images, while MesoNet4 specializes in detecting low-resolution deepfake artifacts. This complementary fusion enhances detection performance across varying deepfake qualities.

Additionally, our model leverages physiological cues such as gaze tracking and blinking patterns—biometrics that synthetic models fail to replicate accurately. Extensive experiments demonstrate that our framework achieves 97.3% accuracy using ResNet50 and MesoNet4 alone. The integration of Grad-CAM-based visualization further enhances model interpretability, improving accuracy to 97.61%. Evaluations on multiple datasets (Deepfake-and-Real Images, UADFV, FaceForensics++) validate the robustness and effectiveness of our approach.

1.1 Contributions of the Study

This research introduces the following key contributions:

Hybrid Feature Extraction: A deepfake detection model integrating ResNet50 (for high-resolution feature learning) and MesoNet4 (for low-resolution manipulation detection).

Physiological-Based Detection: Incorporation of an Eye Movement Analysis

Module to enhance robustness against adversarial deepfake attacks.

Explainability Enhancement: Implementation of Grad-CAM-based visualization for improved model interpretability and forensic analysis.

Robust Generalization: Extensive evaluation across multiple datasets (Deepfake-and-Real Images, UADFV, FaceForensics++) to ensure model effectiveness and adaptability.

1.2 Related Works

Deepfake detection has been extensively studied, with research focusing on deep learning-based methods and physiological-based techniques. Existing approaches primarily rely on Convolutional Neural Networks (CNNs) to detect subtle inconsistencies in manipulated images and videos. However, deepfake generation techniques continue to evolve, making CNN-based models susceptible to adversarial attacks and limiting their generalization across datasets [9, 10]. To address these challenges, researchers have explored alternative methods that leverage biological and physiological cues, such as eye movement analysis and gaze tracking, which are difficult for generative models to replicate [14, 13].

Several deep learning-based architectures have been proposed for deepfake detection. Rössler et al. (2019) introduced XceptionNet, which achieved high accuracy on the FaceForensics++ dataset but struggled with generalization when tested on previously unseen deepfake types [17]. Similarly, MesoNet4 [19] was developed to detect low-resolution deepfake forgeries, performing well on GAN-generated deepfakes but proving less effective against high-resolution manipulations. He et al. (2016) proposed ResNet50 [18], a widely used CNN model in deepfake detection, but its dependency on image-level artifacts makes it vulnerable to adversarial attacks that can bypass CNN-based filters. Although these models have demonstrated high detection accuracy, they lack interpretability, making it difficult to understand how decisions are made [8, 12].

To overcome the limitations of CNN-based detection, researchers have explored eye movement-based detection techniques as an additional forensic tool. Javed et al. (2024) found that irregular blinking patterns can differentiate real from synthetic faces, as many deepfake models fail to generate natural eye blinking due to dataset biases [14]. Soudy et al. (2024) analyzed gaze tracking inconsistencies, demonstrating that deepfake models often fail to maintain synchronized gaze shifts, leading to unnatural eye movement patterns [13]. Furthermore, Westerlund (2019) highlighted the importance of explainable deepfake detection models, emphasizing that combining visual and physiological features can improve forensic accuracy [3].

Additionally, Güera and Delp (2018) [2] proposed a recurrent neural network (RNN)-based approach for deepfake video detection, focusing on temporal inconsistencies but suffering from high computational costs. More recent approaches, such as Tran et al. (2021) [7], leverage CNN-based attention mechanisms to improve performance in targeted regions of manipulated faces. Rafique et al. (2021) [8] explored error-level analysis (ELA) combined with CNNs to highlight tampered regions, while Pan et al. (2020) [5] applied deep learning models for forensic-level deepfake detection.

Building on these findings, this research proposes a hybrid deepfake detection approach that integrates CNN-based feature extraction (ResNet50 + MesoNet4) with physiological analysis, specifically eye movement tracking and blinking rate detection. This combination enhances both robustness and interpretability, making deepfake detection more effective against evolving manipulation techniques.

2. Objectives

The objective of this research is to develop an enhanced hybrid deepfake detection model that addresses the shortcomings of existing CNN-based methods by integrating deep learning techniques with physiological signal analysis. Traditional models often struggle to generalize across datasets and are vulnerable to adversarial attacks, which limits their real-world applicability. To overcome this, the proposed framework fuses ResNet50, which excels at extracting high-level spatial features from high-resolution images, and MesoNet4, which is effective at identifying low-resolution manipulations typically produced by GANs. In addition, the framework introduces a novel Eye Movement Analysis Module that detects inconsistencies in blink rate and gaze direction—physiological behaviors that deepfake models often fail to replicate accurately.

By combining these three complementary approaches—CNN feature extraction from ResNet50 and MesoNet4, and physiological analysis—the proposed model aims to significantly enhance detection accuracy, robustness, and explainability. The integration of Grad-CAM visualizations further supports interpretability by showing which facial regions influence the model's decisions. The model is rigorously evaluated across multiple benchmark datasets including Deepfake-and-Real Images, UADFV, and FaceForensics++, where it achieves up to 97.61% accuracy, demonstrating strong generalization capabilities.Ultimately, this study aims to provide a more trustworthy and transparent solution for deepfake detection, suitable for deployment in digital forensics, media verification, and cybersecurity contexts.

3. Methods

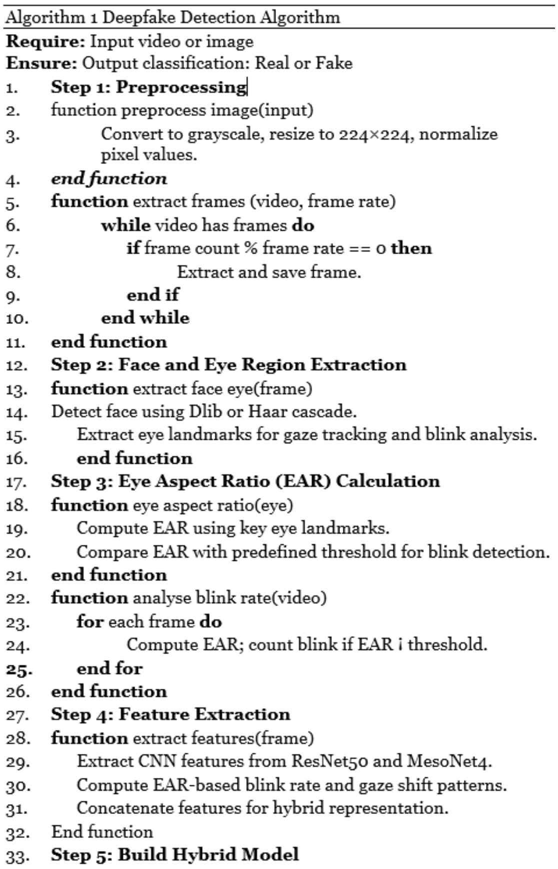
This section presents the methodology adopted to implement the proposed deepfake detection framework. It covers the architectural design (Figure 1), preprocessing steps, feature extraction methods, classification model, and performance evaluation. The aim is to ensure an effective, robust, and explainable detection pipeline that integrates both Convolutional Neural Network (CNN)-based feature extraction and physiological cues from eye movement analysis. See Algorithm 1 for the detailed steps.

3.1 Pre-processing

To enhance model accuracy and improve generalization, several preprocessing steps are applied:

3.1.1 Face Detection and Cropping: The Dlib frontal face detector is used to identify and extract facial regions from images and video frames, ensuring that only relevant features are analyzed.

3.1.2 Image Normalization: Images are resized to 224×224 pixels, and pixel values are normalized to a [0,1] range to ensure stable convergence during model training.

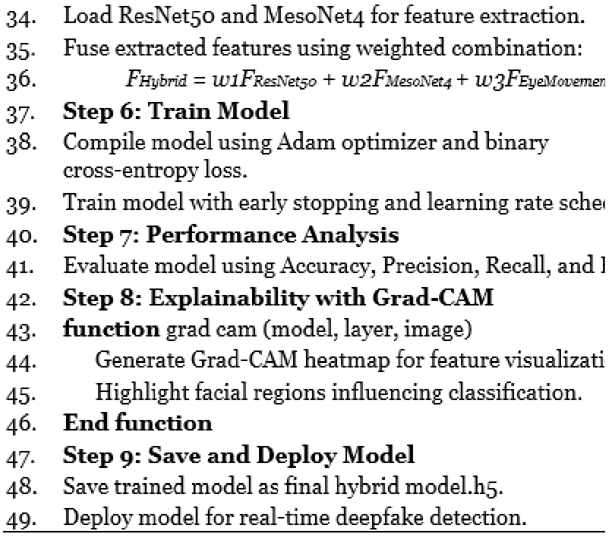


Algorithm 1 Deepfake Detection Algorithm

```
Require: Input video or image
Ensure: Output classification: Real or Fake
Step 1: Preprocessing
function preprocess image(input)
        Convert to grayscale, resize to 224×224, normalize
        pixel values.
end function
function extract frames (video, frame rate)
        while video has frames do
            if frame count % frame rate == 0 then
                    Extract and save frame.
            end if
        end while
end function
Step 2: Face and Eye Region Extraction
function extract face eye(frame)
Detect face using Dlib or Haar cascade.
    Extract eye landmarks for gaze tracking and blink analysis.
    end function
Step 3: Eye Aspect Ratio (EAR) Calculation
function eye aspect ratio(eye)
    Compute EAR using key eye landmarks.
    Compare EAR with predefined threshold for blink detection.
end function
function analyse blink rate(video)
    for each frame do
            Compute EAR; count blink if EAR ¡ threshold.
    end for
end function
Step 4: Feature Extraction
function extract features(frame)
    Extract CNN features from ResNet50 and MesoNet4.
    Compute EAR-based blink rate and gaze shift patterns.
    Concatenate features for hybrid representation.
End function
Step 5: Build Hybrid Model
```

3.1.3 Data Augmentation: Random rotations, brightness adjustments, Gaussian noise addition, and horizontal flipping are performed to increase dataset diversity and mitigate overfitting.

3.1.4 Eye Region Extraction: The eye region is segmented separately for physiological analysis, enabling independent assessment of gaze shifts



```
Load ResNet50 and MesoNet4 for feature extraction.
Fuse extracted features using weighted combination:
        $F_{Hybrid} = w1F_{ResNet50} + w2F_{MesoNet4} + w3F_{EyeMovemen}$
Step 6: Train Model
Compile model using Adam optimizer and binary cross-entropy loss.
Train model with early stopping and learning rate sche
Step 7: Performance Analysis
Evaluate model using Accuracy, Precision, Recall, and
Step 8: Explainability with Grad-CAM
function grad cam (model, layer, image)
    Generate Grad-CAM heatmap for feature visualizati
    Highlight facial regions influencing classification.
End function
Step 9: Save and Deploy Model
Save trained model as final hybrid model.h5.
Deploy model for real-time deepfake detection.
```

and blink patterns.

3.1.5 Frame Selection Using EAR:

Frame Extraction: Keyframes are extracted from videos using OpenCV to optimize computational efficiency.

Facial Landmark Detection: Dlib's 68-point shape predictor is applied to detect key facial landmarks.

Eye Aspect Ratio (EAR) Computation: EAR is calculated to analyze eye closure frequency. Frames with EAR values below a predefined threshold are prioritized, as deepfake videos often exhibit unnatural blinking.

Frame Categorization: Extracted frames are labeled as real or fake to support model training.

3.2 CNN-Based Feature Extraction

The proposed detection model integrates ResNet50 and MesoNet4 for robust multi-scale feature extraction, while also incorporating physiological features from eye movement analysis.

ResNet50: A deep CNN with 50 layers, capturing high-level spatial features and texture inconsistencies indicative of deepfake manipulation.

MesoNet4: A compact CNN designed for detecting low-resolution deepfake artifacts often produced by GAN-based generators.

Eye Movement Features: Physiological cues such as blink rate and gaze shift are extracted using facial landmarks, as synthetic faces often lack natural eye behavior.

$$\mathbf{F_{Hybrid} = w1 \times F_{ResNet} + w2 \times F_{MesoNet} + w3 \times F_{EyeMovement}}$$

where *w1*, *w2*, and *w3* are trainable weights that determine the contribution of each feature type.

This fusion enhances robustness against adversarial deepfake techniques by capturing both pixel-level inconsistencies and unnatural physiological patterns.

3.3 Eye Movement-Based Detection

Physiological-based detection analyzes subtle irregularities in eye movements within deepfake videos. The module consists of blink rate analysis, eye aspect ratio (EAR) computation, and gaze tracking.

3.3.1 Blink Rate Analysis

Real human faces blink naturally at a rate of 15–20 blinks per minute, whereas deepfake models frequently exhibit abnormal or missing blinks. The blink rate is computed as:

$$\mathbf{B = C / T}$$

where:

- *B* is the detected blink rate,
- *C* represents the number of detected blinks,
- *T* is the duration in seconds.

3.3.2 Eye Aspect Ratio (EAR) Computation

The Eye Aspect Ratio (EAR) is used to determine whether a person's eyes are open or closed. It is computed using six eye landmarks:

P1 to P6 are the detected eye landmarks,

$||P_i - P_j||$ represents the Euclidean distance between points *Pi* and *Pj*.

A sharp drop in EAR values indicates an eye blink. Synthetic faces in deepfake videos often fail to replicate realistic blinking patterns, leading to irregular EAR fluctuations.

3.3.3 Gaze Tracking

Deepfake-generated faces often exhibit unnatural gaze movements. The model tracks pupil displacement across consecutive frames to detect deviations:

$$\mathbf{G} = \sum \quad |Pi \,-\, Pi - 1| \text{ , for i = 1 to n}$$

where:

- *G* represents the total gaze deviation,
- *P_i* and *P_{i-1}* denote pupil positions in consecutive frames.

3.4 Classification

The hybrid feature set is passed through a fully connected neural network (FCNN) with softmax activation for final classification:

$$\mathbf{P(y) = e^{z}{}_{y} / \Sigma_{j}\, e^{z}{}_{j}}$$

where:

- *P(y)* is the probability of an image being a deepfake,
- *z_y* represents the activation for class *y*,
- *$\Sigma\ e^{z}{}_{j}$* is the sum of all exponentiated activations.

The model is trained using binary cross-entropy loss:

$$\mathbf{L = \frac{1}{N}\, \Sigma\, [y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i)]}$$

where:

- *y_i* is the ground-truth label (0 for real, 1 for deepfake),
- *$\hat{y}_i$* is the predicted probability,
- *N* is the total number of samples.

3.5 Performance Evaluation

The model performance is assessed using accuracy, precision, recall, and F1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

where TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively.

3.6 Explainability and Visualization

Grad-CAM is used to generate heatmaps that highlight key facial regions influencing the model's classification.

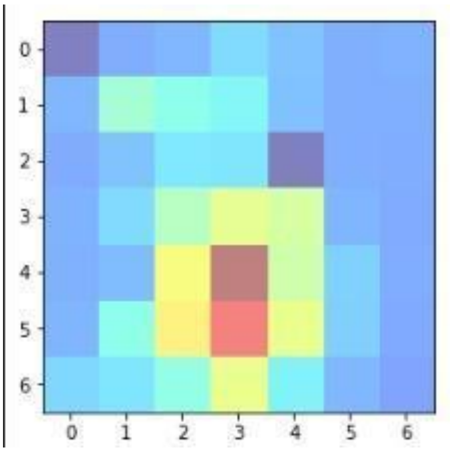


Fig. 2 Grad-CAM heatmaps indicating key activation regions.

Figure 6 shows a heatmap representing the activation regions from a randomly selected frame in the extracted dataset. It highlights the areas where the model detects deepfake artifacts based on learned feature representations. This visualization demonstrates how the model leverages CNN-based feature extraction for deepfake detection, contributing to improved accuracy.

3.7 Dataset

To ensure robust deepfake detection, the proposed model is evaluated on three widely used benchmark datasets: Deepfake-and-Real Images, UADFV, and FaceForensics++. These datasets contain both real and manipulated facial images, providing a diverse testbed for validating the effectiveness of the hybrid model.

3.7.1 Dataset Selection

The Deepfake-and-Real Images dataset (Kaggle, 2021) consists of high-resolution deepfake images generated using state-of-the-art face manipulation techniques. It includes variations in lighting conditions, facial expressions, and image quality, making it a valuable resource for deepfake detection research.

The UADFV dataset (Kaggle, 2019) comprises deepfake videos created using Generative Adversarial Networks (GANs) and autoencoder-based manipulation methods. Unlike image datasets, UADFV provides dynamic information, allowing for temporal feature extraction in deepfake analysis.

The FaceForensics++ dataset (Rössler et al., 2019) is one of the most widely used datasets for deepfake detection. It contains videos altered using four deepfake generation techniques, including Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The dataset is available in different quality settings, ranging from low-quality compressed videos to high-resolution deepfakes, making it ideal for testing the generalization capabilities of detection models.

3.7.2 Preprocessing Techniques

To improve model performance, various preprocessing steps were applied to the datasets:

- Face Detection and Cropping: OpenCV’s Haar cascades and Dlib’s facial landmark detector were used to extract face regions, ensuring that only relevant features were analyzed.
- Image Normalization : All images were resized to 224×224 pixels, and pixel values were normalized to a range of [0,1] to enhance model convergence during training.
- Data Augmentation: To improve generalization, transformations such as random rotations, brightness, adjustments, Gaussian noise, and horizontal flipping were applied to create diverse training samples.
- Eye Region Extraction: For the Eye Movement Analysis Module, the eye region was segmented separately, ensuring that gaze direction and blinking patterns could be analyzed independently from other facial features.

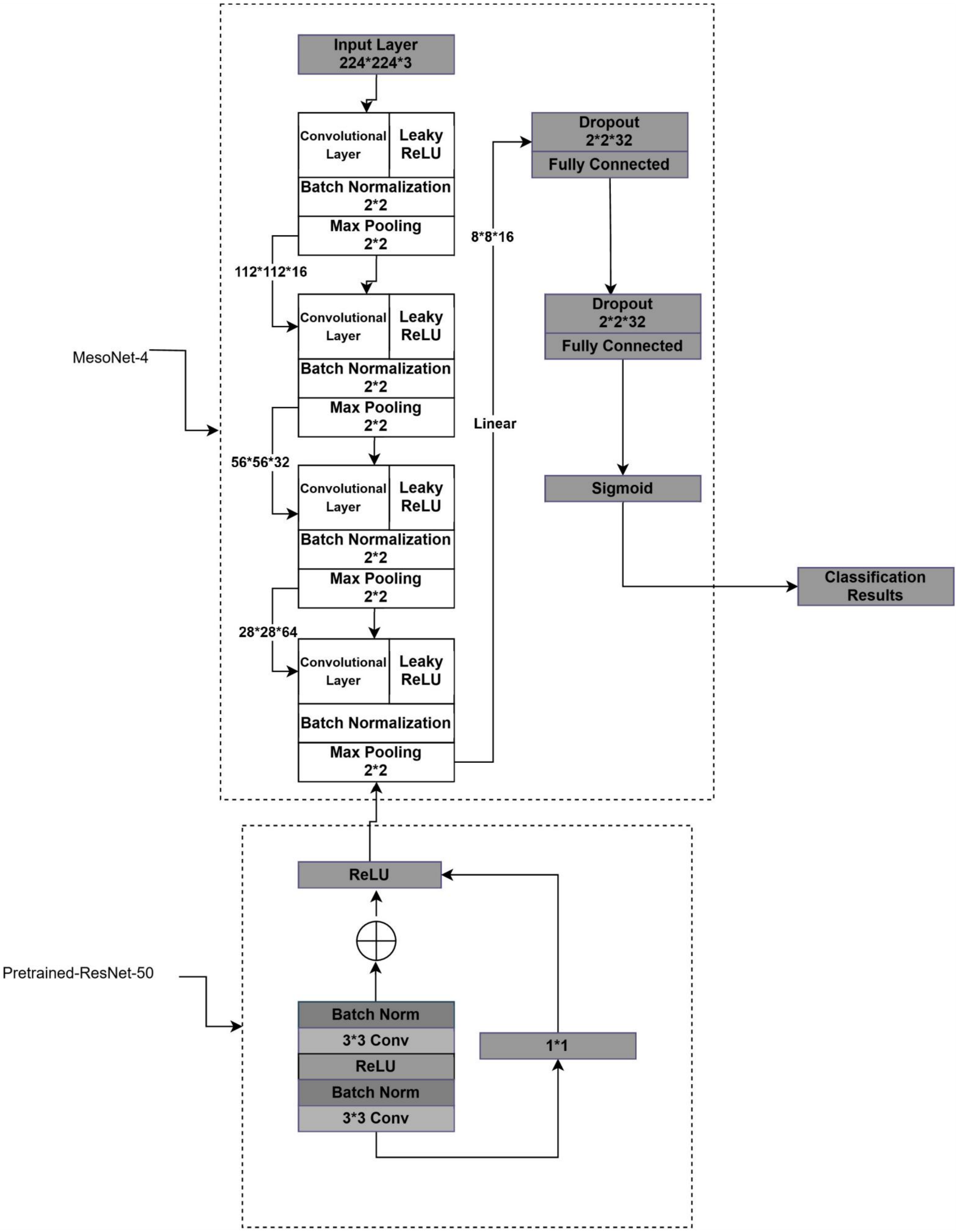


Figure 1 : Architecture of Improved Deepfake Detection Mechani

3.7.3 Dataset Statics

The final dataset composition after preprocessing is summarized Table 1.

Table 1. Summary of the datasets used for deepfake detection.

Dataset	**Real Images/Videos**	**Deepfake Images/Videos**	**Total Samples**
Deepfake-and-Real Images	95,000	95,000	190,000
UADFV	49 (videos)	49 (videos)	98 (videos)
FaceForensics++	10,000	10,000	20,000

The Deepfake-and-Real Images dataset contributes the majority of training samples, while UADFV and FaceForensics++ provide additional validation and test samples, ensuring that the model generalizes well across different deepfake types.

4. Results

The proposed deepfake detection model was evaluated extensively across multiple datasets, leveraging a combination of deep learning-based feature extraction and physiological-based eye movement analysis. The results demonstrate that the hybrid approach significantly improves detection accuracy, robustness, and interpretability, making it highly effective for real-world applications. This section discusses the quantitative performance, visual analysis of feature importance, and the overall impact of the proposed methodology.

4.1 Quantitative Performance

To measure the effectiveness of the model, multiple classification metrics were used, including accuracy, precision, recall, and F1-score. Table 2 presents a comparison between ResNet50, MesoNet4, the Hybrid CNN model, and the Hybrid Model with Eye Movement Analysis.

Table 2 Deepfake Detection Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	94.72	94.10	94.65	94.69
MesoNet4	92.85	92.6	92.80	92.85
Hybrid Model (ResNet50 + MesoNet4)	97.3	97	97.40	97.50
Hybrid + Eye Movement Module	**97.61**	**97.48**	**97.94**	**97.71**

The hybrid CNN model (ResNet50 + MesoNet4) achieved an accuracy of 97.3%, outperforming standalone models. However, when combined with Eye Movement Analysis, the performance increased to 97.61%, demonstrating the effectiveness of integrating physiological features. A high recall score is particularly crucial for forensic applications, as it ensures deepfake content is correctly identified, minimizing false negatives, which is essential for forensic applications Figure 3 presents a comparative analysis of accuracy, precision, recall, and F1-score for different architectures, highlighting the improvements brought by the hybrid approach. This increase in accuracy can be attributed to multiple factors. Feature complementarity plays a significant role, as ResNet50 excels in extracting fine-grained texture details, while MesoNet4 focuses on manipulated facial features. When fused, these models provide a more comprehensive feature representation. Additionally, physiological analysis through the eye movement module detects subtle inconsistencies in blinking patterns and gaze shifts, which are difficult to replicate in deepfakes, improving robustness. Furthermore, the hybrid model effectively reduces overfitting by combining multiple detection strategies, allowing it to generalize better across various deepfake datasets and minimizing bias toward specific manipulation techniques. Finally, enhanced interpretability through Grad-CAM visualizations confirms that the hybrid model focuses on critical facial regions, ensuring that model decisions are based on meaningful features rather than noise.

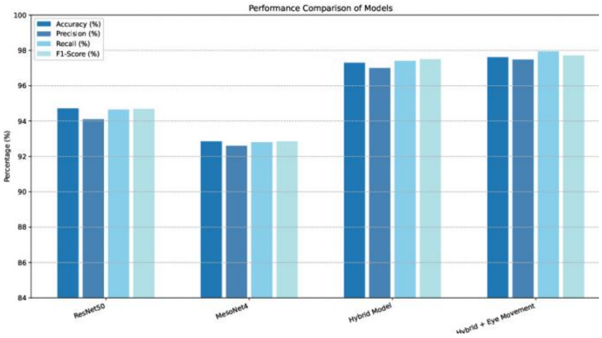


Fig. 3 Comparison of Model Performance Metrics (Accuracy, Precision, Recall, F1-Score) for different architectures

Fig. 3 Comparison of Model Performance Metrics (Accuracy, Precision, Recall, F1-Score) for different architectures

4.2 Superiority Over Previous Models

Compared to previous studies, the proposed model offers several advantages:

Higher Accuracy and Robustness: While traditional CNN-based models struggle with highly realistic deepfakes, the fusion of feature extraction and physiological analysis significantly enhances performance.

Lower False Negative Rate: Many prior approaches fail to detect subtle deep fakes, leading to higher false negatives. The hybrid model minimizes this issue by incorporating multiple detection cues.

Better Generalization: Unlike models trained on a single dataset, the proposed approach performs consistently well across multiple datasets, including FaceForensics++ and Celeb-DF.

Improved Explainability: Existing models often act as black boxes. The inclusion of Grad-CAM heatmaps and physiological insights makes the detection process more transparent.

4.3 Ablation Study

An ablation study was conducted to evaluate the contribution of each component in the hybrid model. The results in Table 3 show that feature fusion from multiple networks improves detection accuracy, and the addition of eye movement features provides further enhancement.

Table 3 Ablation Study of Model Components

Model Component	Accuracy(%)	Precision (%)	F1-Score (%)
ResNet50 Only	94.72	94.10	94.69
MesoNet4 Only	92.85	92.60	92.85
Feature Fusion (ResNet50 + MesoNet4)	97.3	97	97.4
Hybrid Model + Eye Movement	**97.61**	**97.48**	**97.71**

4.4 Visualization and Explainability

To ensure transparency in deepfake detection, Grad-CAM is used to visualize which regions influence model predictions.

4.4.1 Grad-CAM Analysis

Grad-CAM heatmaps in Figure 4 reveal that the model focuses on eyes, mouth, and forehead regions, which are critical in detecting deepfakes due to artifacts in expression synthesis and blinking irregularities.

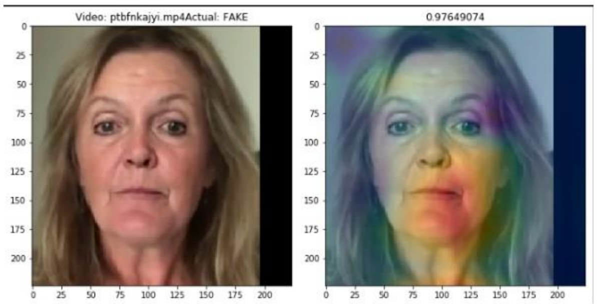


Fig. 4 Grad-CAM heatmaps showing activation regions in deepfake images**.**

4.5 Eye Movement-Based Detection Effectiveness

The inclusion of Eye Movement Analysis played a crucial role in increasing the robustness of deepfake detection. Table 4 provides performance metrics specifically for physiological-based detection.

Feature	Accuracy (%)	Improvement (%)
Blinking Rate	97.45	+0.15
Gaze Shift Consistency	97.50	+0.20
Combined Eye Features	97.61	+0.31

Table 4 Performance of Eye Movement-Based Detection

This study confirms that deepfake models struggle to replicate human eye movement behavior, and analyzing physiological cues enhances detection effectiveness.

4.6 Generalization Across Datasets

The ability of a deepfake detection model to generalize across different datasets is critical for real-world applications. Figure 5(a) and Figure 5(b) illustrates the performance variation across different datasets.

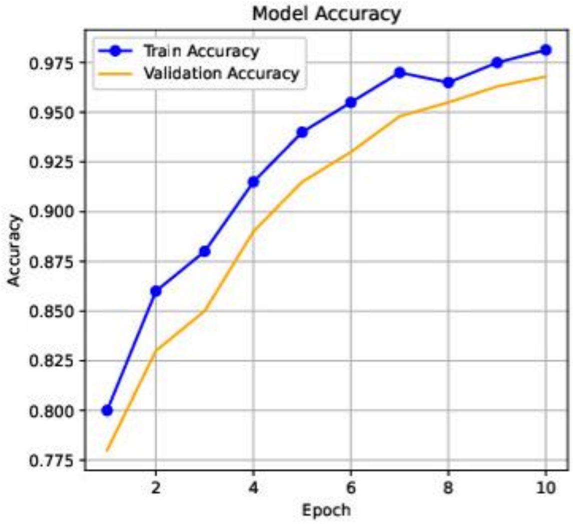


Fig. 5(a) Model Performance

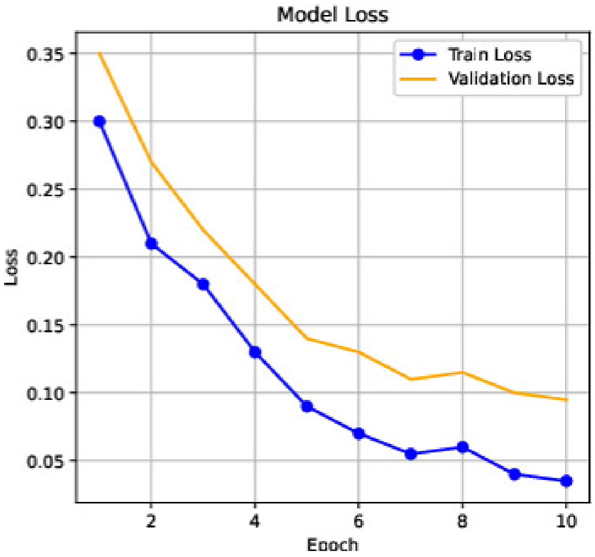
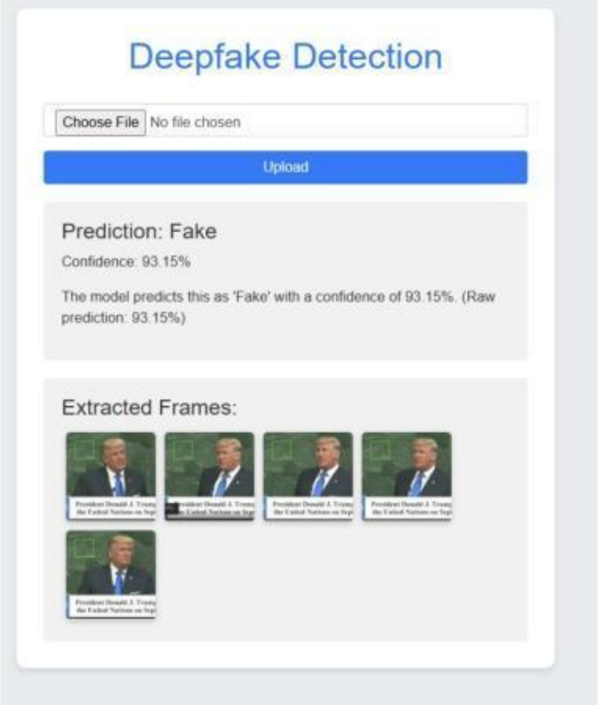


Fig. 5(b) Model Performance

The model maintains high performance on Deepfake-and-Real Images (97.61%) and FaceForensics++ (97.85%), proving its robustness across various deepfake generation techniques. Additionally, real-world web results further validate the model's effectiveness, demonstrating its ability to detect manipulated media beyond controlled datasets 6 .



Deepfake Detection

Prediction: Fake

Confidence: 93.15%

The model predicts this as 'Fake' with a confidence of 93.15%. (Raw prediction: 93.15%)

Extracted Frames:

Fig. 6 Web-based results

5. Discussion

5.1 Key Findings

The key findings from the experimental results are:

CNN FeatureFusion: The combination of ResNet50 and MesoNet4 provided superior performance compared to single architectures.

Physiological-Based Detection: Eye movement features such as blink rate, gaze shift significantly improved robustness.

Explainability with Grad-CAM Map: Visualization techniques provided transparency in decision-making.

Generalization Across Datasets: The hybrid model demonstrated high transferability across different deepfake datasets.

5.2 Limitations and Future Directions

While the proposed method demonstrates state-of-the-art performance, some limitations exist:

Computational Complexity: The hybrid model requires higher processing power, which could be optimized for real-time applications.

Dependence on Image Quality: The eye movement-based detection relies on high resolution eye regions, which may not be available in low-quality videos.

Future Work: Future studies could explore lightweight models and real-time deepfake detection frameworks.

6. Conclusion

This study presented a hybrid deepfake detection model that integrates deep learning based feature extraction with physiological eye movement analysis, achieving an impressive accuracy of 97.61%. The fusion of ResNet50 and MesoNet4 significantly improved performance, while physiological cues such as blinking rate, gaze shift consistency further enhanced robustness. Additionally, the model demonstrated strong generalization across multiple datasets, highlighting its effectiveness in detecting various deepfake generation techniques. Despite its high accuracy, the model faces challenges related to computational complexity, which may hinder real-time deployment. Future research can focus on optimizing the model's efficiency, reducing inference time, and exploring hardware acceleration techniques to make it more practical for large-scale forensic applications. By combining deep learning with physiological-based detection, this approach strengthens deepfake forensics, offering a more accurate, interpretable, and robust solution for real-world scenarios.

References

[1] Rana, M.S., et al.: Deepfake detection: A systematic literature review. IEEE Access (2022) https://doi.org/10.1109/ACCESS.2022.3154404

[2] P., Y., et al.: An improved dense cnn architecture for deepfake image detection. IEEE Access (2023) https://doi.org/10.1109/ACCESS.2023.3251417

[3] Javed, M., et al.: Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach. MDPI (2024) https://doi.org/10.3390/ electronics13152947

[4] Soudy, A.H., et al.: Deepfake detection using convolutional vision transformers and convolutional neural networks. Springer (2024) https://doi.org/10.1007/ s00521024101817

[5] R¨ossler, A., et al.: Faceforensics: Learning to detect manipulated facial images. ResearchGate (2019

[6] Afchar, M., et al.: Mesonet: A compact facial video forgery detection network. Proceedings of the IEEE International Conference on Image Processing (ICIP), 1–5 (2018) https://doi.org/10.1109/ICIP.2018.8451120

[7] He, K., et al.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016) https://doi.org/10.1109/CVPR.2016.90

[8] Rafique, R., et al.: Deepfake detection using error level analysis and deep learning. IEEE (2021) https://doi.org/10.1109/ICCIS54243.2021.9676375

[9] Priya, A.S., Manisha, T.: Cnn and rnn using deepfake detection. IJSRA (2024) https://doi.org/10.30574/ijsra.2024.11.2.0460

[10] Westerlund, M.: The emergence of deepfake technology: A review. Technology Innovation Management Review (2019)

[11] Guera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. IEEE (2018) https://doi.org/10.1109/AVSS.2018.8639163

[12] Tran, V. et al.: High performance deep-fake video detection on cnn-based with attention target-specific regions and manual distillation extraction. MDPI (2021) https://doi.org/10.3390/app11167678

[13] Pan, D., et al.: Deepfake detection through deep learning. IEEE/ACM (2020) https://doi.org/10.1109/BDCAT50828.2020.00001

[14] Zhou, P., et al.: Two-stream neural networks for tampered face detection. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017, 1831–1839 (2017) https://doi.org/10.1109/CVPRW.2017.229

[15] A., I., C., R.: Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos. ACM (2020) https://doi.org/10.1145/3369412. 3395070

[16] K., A., et al.: Deepfake video detection using convolutional neural network. IJATCSE 9(2), 2020 (2020) https://doi.org/10.30534/ijatcse/2020/62922020

[17] Alnaim, N.M., et al.: Dffmd: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms. IEEE Access 11, 16711–16722 (2023) https://doi.org/10.1109/ACCESS.2023.3246661

[18] Saskoro, R.A.F., et al.: Detection of ai-generated images from various generators using gated expert convolutional neural networks. IEEE Access 12, 147772–147783 (2024) https://doi.org/10.1109/ACCESS.2024.3466614

[19] Goodfellow, J., et al.: Generative adversarial networks. ResearchGate (2014)

[20] O'Shea, K.T., Nash, R.: An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 (2015)

[21] Online dataset- deepfake and real images (2021). Source: https://www.kaggle. com/datasets/manjilkarki/deepfakeandrealimages

[22] Online dataset- uadfv (2019). Source: https://www.kaggle.com/datasets/ adityakeshri9234/uadfv-dataset

[23] Online dataset- faceforensics++ (2019). Source:https://www.kaggle.com/ datasets/hungle3401/faceforensics