INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, MARCH 2026 | ISSN: 2456-4184 | IJNRD.ORG

Deepfake Face Swap Detection System

Ujwal Bagul[1], Vedant Shimpi[2], Avadhoot Chavan[3], Manish Mhatre[4], Dr. Vishwajit Balaso Gaikwad[5]

[1]Student, [2]Student, [3]Student, [4]Student, [5]Associate Professor
[1]Department of Computer Engineering
[1]Terna Engineering College, Navi Mumbai, India

ABSTRACT

Deepfake face-swapping technologies, powered by advances in artificial intelligence and computer vision, have enabled the creation of highly realistic synthetic media that is increasingly difficult to distinguish from authentic content. While such technologies have legitimate applications in entertainment and media production, their misuse poses significant threats, including misinformation, identity theft, and digital forgery. Existing deepfake detection approaches often rely on static image analysis, lack robustness against emerging manipulation techniques, and fail to incorporate multi-modal signals or explainability. This project proposes an advanced deep learning-based system for detecting face-swapped images and videos using a hybrid architecture that integrates spatial, temporal, and audio-visual analysis. The system employs Convolutional Neural Networks (CNNs) for spatial feature extraction, Recurrent Neural Networks (RNNs) for temporal inconsistency detection, and audio-visual synchronization models to identify lip-sync mismatches. Additionally, explainable artificial intelligence techniques such as Grad-CAM and LIME are incorporated to provide visual justifications for model predictions, enhancing transparency and trust. The proposed system processes both images and videos by performing frame extraction, face alignment, and feature analysis, followed by classification into real or manipulated categories. A web-based interface enables users to upload media and receive detailed reports, including probability scores, detection timestamps, and heatmaps highlighting manipulated regions. Experimental evaluation demonstrates that the system achieves high accuracy and improved robustness compared to existing methods, particularly in handling compressed and real-world media scenarios, as also reflected in the performance metrics summarized in the results section

Keywords **–** Deepfake Detection, Convolutional Neural Networks, Face Swap Detection , Audio Visual Analysis, Explainable AI.

I. INTRODUCTION

In recent years, the rapid advancement of Artificial Intelligence (AI) and Deep Learning (DL) has significantly contributed to the development of sophisticated deepfake technologies capable of generating highly realistic manipulated media [1]. Early deep learning-based approaches focused on detecting pixel-level inconsistencies and visual artifacts in manipulated images; however, these methods were limited to spatial analysis and failed to capture temporal inconsistencies in video data [1]. The availability of large-scale datasets such as eKYC-DF has further supported the training and evaluation of deepfake detection models by providing diverse manipulation scenarios, although such datasets often lack generalization to real-world applications [2]. Subsequent research explored advanced feature fusion and federated learning techniques to improve detection accuracy and data privacy [3], while attention-based models using architectures such as EfficientNet demonstrated improved performance by focusing on critical facial regions like eyes and lips [4]. Additionally, hybrid approaches combining frequency domain analysis with transformer-based models were introduced to capture spectral artifacts that are difficult to manipulate, achieving high robustness at the cost of increased computational complexity [5]. To address the limitations of unimodal systems, cross-modal approaches integrating both audio and visual features have been proposed, enabling detection of lip-sync inconsistencies and facial motion mismatches [6]. Lightweight deep learning models were also developed for deployment on mobile and edge devices, offering faster inference but often sacrificing accuracy and robustness [7]. Despite these advancements, studies have shown that most deepfake detection models struggle to generalize across different datasets and unseen manipulation techniques, highlighting a critical challenge in real-world deployment [8]. Further improvements have been made through video-based models that incorporate temporal analysis using facial landmarks, convolutional operations, and attention mechanisms to capture sequential inconsistencies across frames [9]. Comprehensive surveys have analyzed existing deepfake detection methods and identified key challenges such as dataset bias, lack of robustness, and vulnerability to adversarial attacks [10]. These studies emphasize the growing threat of deepfake technology across multiple domains and highlight the importance of developing multi-modal detection systems that integrate visual, audio, and temporal information for improved accuracy and reliability [11]. Additionally, research has pointed out the lack of unified frameworks capable of handling both image and video data effectively, along with the need for scalable and interpretable detection systems suitable for real-world applications [12]. To address these challenges, this work

proposes a comprehensive deepfake face swap detection system that integrates spatial, temporal, and audio-visual analysis within a unified deep learning framework. The proposed system utilizes CNN-based architectures for extracting spatial features from images and video frames, while Recurrent Neural Networks (RNNs), specifically Bidirectional Long Short-Term Memory (BiLSTM) networks, are employed to capture temporal dependencies across sequential frames. Furthermore, audio-visual synchronization techniques are incorporated to detect inconsistencies between speech and lip movements, enhancing detection accuracy in video-based scenarios. To improve interpretability, Explainable Artificial Intelligence (XAI) methods such as Grad-CAM and LIME are integrated to provide visual explanations of model predictions. The system is implemented as a web-based platform that enables users to upload images or videos and receive detailed analysis results, including classification outputs, confidence scores, and visual explanation maps, making it suitable for real-world applications such as media authentication, cybersecurity, and digital forensics.

II. LITERATURE REVIEW

The rapid advancement of deep learning technologies has significantly influenced the development of deepfake detection systems, aiming to counter the increasing threat of manipulated digital media. Early approaches primarily focused on identifying visual inconsistencies within images using deep learning models. Kim and Cho (2020) [1] proposed a deep learning-based framework that combines content and trace feature extractors to detect pixel-level artifacts introduced during face manipulation. While the model achieved high accuracy, it was limited to spatial analysis and failed to address temporal inconsistencies present in video data, such as unnatural blinking or frame transitions.

The availability of large-scale datasets has further accelerated research in this domain. The eKYC-DF dataset introduced by researchers [2] provided diverse manipulation techniques and facial variations, improving model training and evaluation. However, its application remains restricted to identity verification scenarios, limiting its adaptability to broader real-world deepfake detection tasks. To enhance performance and privacy, feature fusion-based models using federated learning were proposed [3], which combine representations from distributed datasets. Although these methods improve robustness, they are computationally expensive and lack integration of temporal and audio-visual features.

Recent advancements have focused on improving feature extraction using attention mechanisms and advanced architectures. An EfficientNet-based attention model proposed by researchers [4] demonstrated improved detection accuracy by focusing on subtle facial regions such as eyes and lips. Similarly, hybrid approaches combining frequency domain analysis with Vision Transformers [5] have been developed to capture spectral artifacts that are difficult to manipulate. Despite achieving high accuracy, these methods require significant computational resources, making them less suitable for real-time applications.

To overcome the limitations of unimodal systems, cross-modal approaches integrating both audio and visual information have gained attention. A cross-modal attention-based model [6] analyzed lip-sync inconsistencies and facial motion patterns to improve detection accuracy in video-based scenarios. However, such models depend heavily on synchronized and high-quality audio-visual data, limiting their effectiveness in noisy or real-world environments. Lightweight deep learning models have also been proposed for deployment on mobile and edge devices [7], offering faster inference and reduced latency, but often at the cost of reduced accuracy and inability to detect complex manipulations.

Another major challenge in deepfake detection is the lack of generalization across datasets. Studies have shown that most models perform well on specific datasets but fail to detect unseen deepfake techniques [8], highlighting the need for more robust and adaptable systems. To address temporal inconsistencies, video-based models incorporating facial landmark extraction, convolutional operations, and self-attention mechanisms have been developed [9]. These models improve detection by analyzing sequential patterns across frames but require high computational resources and large datasets.

Comprehensive surveys in this domain [10] have identified key challenges such as dataset bias, lack of robustness, and vulnerability to adversarial attacks. Further studies [11] emphasize the growing impact of deepfake technology across visual and audio domains, highlighting the importance of multimodal detection systems. Additionally, research [12] has pointed out the absence of unified frameworks capable of handling both image and video data effectively, along with the lack of scalable and interpretable solutions for real-world deployment.

III. METHODOLOGY

The proposed Deepfake Detection System adopts a modular and systematic architecture that integrates data preprocessing, multi-modal deep learning-based analysis, explainable artificial intelligence, backend deployment, and automated report generation. The system is designed to process both images and videos by combining spatial, temporal, and audio-visual features, thereby improving detection accuracy, robustness, and interpretability in real-world scenarios.

A block diagram of the proposed system is shown in Figure 1, illustrating the overall workflow from input acquisition to final output generation.

A. Overall System Workflow

The overall workflow of the proposed system is illustrated in **Figure 1**. The system accepts an input in the form of an image or video, which first undergoes preprocessing to standardize the data, including resizing, normalization, frame extraction (for videos), and audio separation. This step ensures consistency and prepares the data for efficient model processing. Following preprocessing, relevant features are extracted from both visual and audio components. The extracted features are then analyzed through multiple specialized branches, including spatial artifact detection for identifying visual inconsistencies, temporal consistency analysis for capturing frame-level irregularities, and audio-visual synchronization to detect mismatches between speech

and lip movements. Each branch focuses on a specific type of deepfake manipulation, enabling a more comprehensive analysis. The outputs from these branches are subsequently fused to generate a final prediction indicating whether the input is real or fake. To enhance transparency, explainability techniques such as saliency mapping are applied to highlight important regions influencing the model's decision. Finally, the system produces a detailed report containing the classification result, confidence score, and visual explanations, making the solution suitable for real-world applications such as digital forensics and media verification.

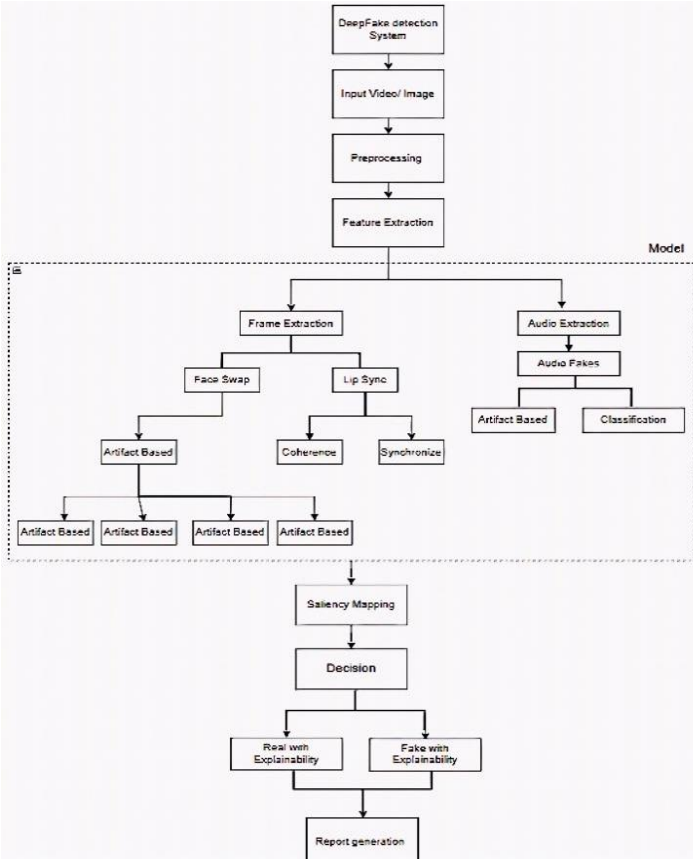

Figure 1. Functional Block Diagram of Deepfake Detection Model

B. Data Preprocessing

The input media, which may be an image or video, undergoes preprocessing to standardize and prepare the data for model input. For video data, frames are extracted using FFmpeg at fixed intervals, and the audio stream is separated for further analysis. Face detection and alignment techniques are applied to focus on relevant facial regions and eliminate background noise.

All frames are resized and normalized to maintain consistency across inputs. Data augmentation techniques such as flipping, rotation, and brightness adjustment are applied to improve model robustness against real-world variations.

C. Multimodal System Architecture

The architecture of the proposed system is shown in **Figure 2**, which consists of two primary branches: the audio branch and the visual branch. The visual branch focuses on extracting frame-level features from images or video frames to detect spatial artifacts, facial inconsistencies, and temporal irregularities. It includes modules for face swap cue detection, artifact analysis, lip region feature extraction, and temporal coherence modeling to capture frame-to-frame variations. In parallel, the audio branch processes the extracted audio signal to identify anomalies such as synthetic speech patterns and inconsistencies in audio characteristics. Both branches operate independently to capture modality-specific features and are later integrated through a fusion layer, which combines the complementary information from audio and visual streams. This multimodal fusion enhances the overall detection performance by enabling the system to analyze deepfake content more comprehensively and robustly across different types of manipulations.

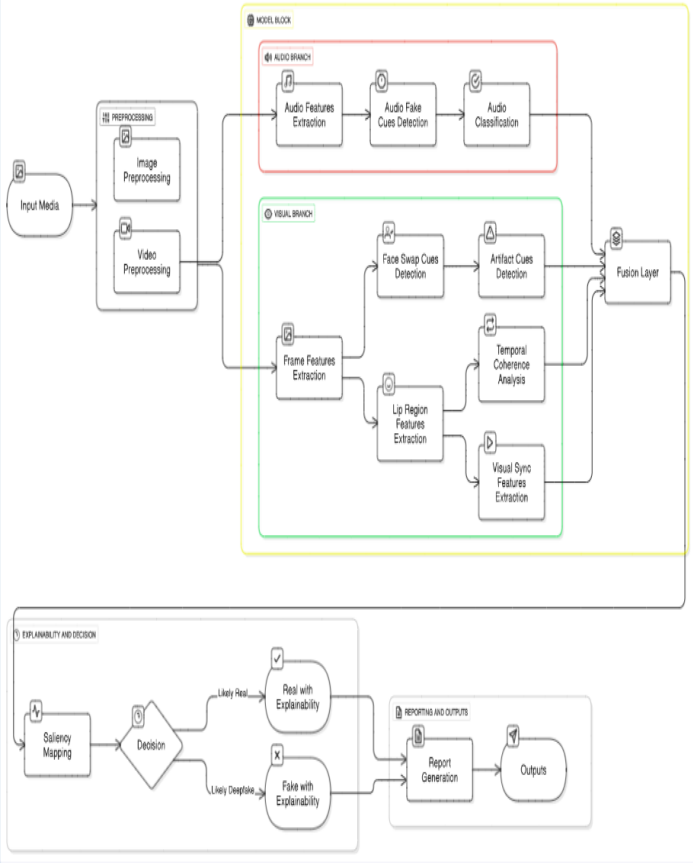

Figure 2. Architecture of Multimodal Deepfake Detection System

1. Visual Branch

The visual branch extracts frame-level features and analyzes spatial and temporal inconsistencies. It includes:

- Face swap cue detection
- Artifact detection using convolutional neural networks
- Lip region feature extraction
- Temporal coherence analysis
- Visual synchronization feature extraction

These components help detect anomalies such as unnatural facial textures, inconsistent motion, and lip-sync mismatches.

2. Audio Branch

The audio branch processes the extracted audio signal to identify inconsistencies such as unnatural speech patterns or synthetic audio artifacts. It includes:

- Audio feature extraction
- Detection of audio manipulation cues
- Audio classification

3. Multimodal Fusion

The outputs from both branches are combined using a fusion layer to improve detection accuracy. This multimodal approach enables the system to leverage complementary information from both visual and audio domains.

D. Deep Learning Model

The core deep learning model used for classification is illustrated in **Figure 3**.

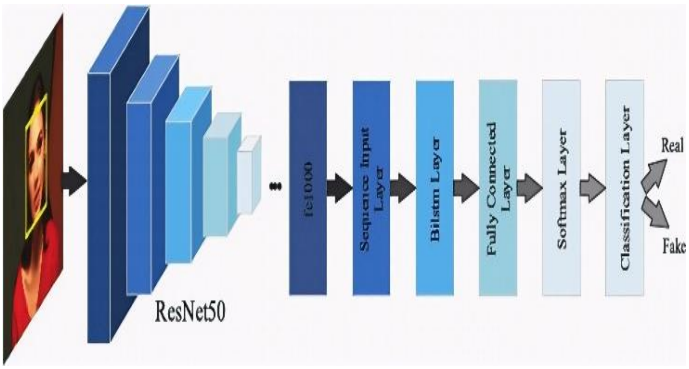

Figure 3. ResNet 50 Model Diagram

The proposed deepfake detection system employs a supervised binary image classification approach based on transfer learning, utilizing a ResNet50 convolutional neural network pretrained on the ImageNet dataset. The pretrained backbone acts as a fixed feature extractor, leveraging learned representations such as edges, textures, and structural patterns to detect subtle deepfake artifacts. The dataset is divided into training, validation, and testing sets, with all images resized to 180×180 pixels and processed in mini-batches of size 32. The ResNet50 backbone is frozen (trainable = False) to retain pretrained knowledge and prevent overfitting. The network utilizes residual learning, where each block learns a residual mapping defined as $F(x) = H(x) - x$, and the final output becomes $y = F(x) + x$, enabling efficient gradient propagation and mitigating the vanishing gradient problem. The extracted feature maps are reduced using Global Average Pooling (GAP), computed as $GAP = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_{ij}$, converting high-dimensional feature maps into a compact vector representation. This vector is passed through a fully connected dense layer of 384 neurons, defined by the transformation $z = W \cdot x + b$, followed by a non-linear activation $a = \max(0, z)$using the ReLU function. To prevent overfitting, a dropout layer with probability $p = 0.5$is applied, where neuron outputs are modified as $\tilde{y}_i = \frac{y_i \cdot Bernoulli(1-p)}{1-p}$. The final output layer uses a sigmoid activation function $\sigma(z) = \frac{1}{1+e^{-z}}$, producing a probability score for binary classification. The model is trained using Binary Cross-Entropy loss, defined as $L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$, which penalizes incorrect predictions. Optimization is performed using the Adam optimizer, where parameter updates are computed as $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$, $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$, followed by bias correction $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$, $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$, and final update rule $\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$. Training is conducted for up to 10 epochs with Early Stopping and Model Checkpoint mechanisms to ensure optimal performance and prevent overfitting. The overall architecture efficiently combines pretrained feature extraction with a lightweight classification head, achieving a balance between computational efficiency and high detection accuracy.

E. Attention-Based Feature Enhancement

To further improve feature extraction, an attention-based convolutional module is incorporated, as shown in **Figure 4**.

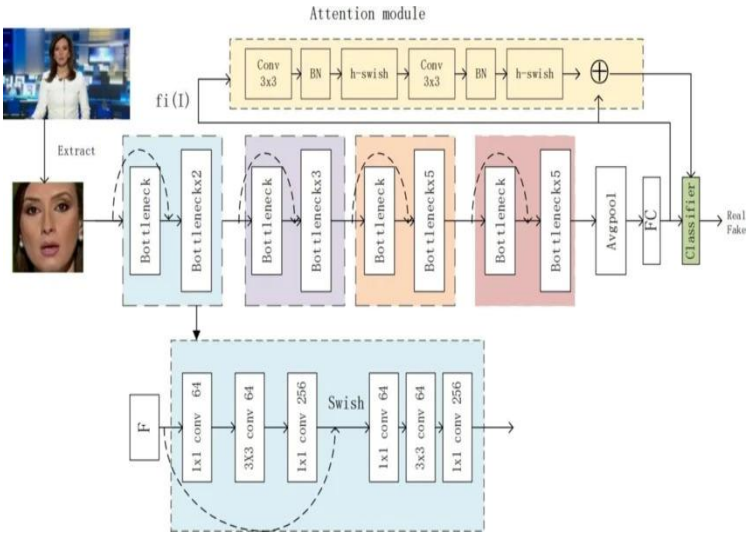

Figure 4. Attention Model Diagram

To improve feature extraction, an attention-based convolutional module is integrated into the model to highlight subtle deepfake artifacts. It applies convolution, normalization, and activation to emphasize important facial regions such as eyes and lips while suppressing irrelevant background features. The refined feature map is given by $F_{attn} = F \odot \sigma(W * F + b)$, where $F$is the input feature map and $\sigma$is the sigmoid function. This mechanism enables the model to focus on discriminative regions, improving detection accuracy and robustness.

F. Multimodal Deepfake Detection Architecture

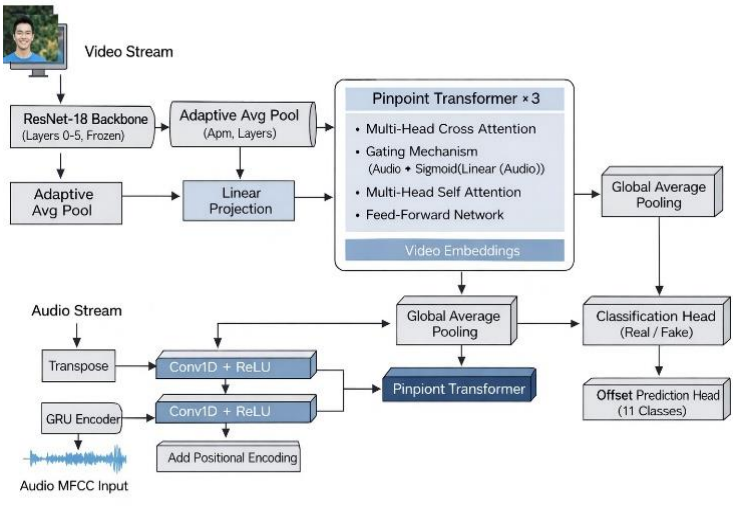

Figure 5. ResNet18 Architecture

The proposed ResNet18-based deepfake detection model formulates image-level classification as a spatial artifact detection problem by leveraging transfer learning and residual feature extraction. The system processes input images Ithrough a ResNet18 backbone, where initial convolutional layers extract low-level features such as edges and textures, while deeper residual blocks capture higher-level semantic inconsistencies introduced by deepfake generation. The defining characteristic of ResNet18 lies in its residual learning framework, where each block learns a residual mapping defined as $F(x) = H(x) - x$, and the final output of the block is computed as $y = F(x) + x$, enabling direct gradient flow and mitigating the vanishing gradient problem. The extracted feature maps $F_{map} \in \mathbb{R}^{H\times W\times C}$are then compressed using Global Average Pooling (GAP), computed as $X = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} f_{ij}$, which reduces spatial dimensions and produces a compact feature vector representation. This feature vector is passed through a fully connected projection layer defined as $z = W \cdot X + b$, followed by a non-linear activation $a = \max(0, z)$using the ReLU function to introduce non-linearity and sparsity in the learned representations. To enhance generalization and prevent overfitting, a dropout mechanism is applied where neuron activations are modified as $\tilde{a}_i = \frac{a_i \cdot Bernoulli(1-p)}{1-p}$, with dropout probability $p = 0.5$. The final classification layer utilizes a sigmoid activation function defined as $\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$, which maps the output to a probability score representing the likelihood of the input being a deepfake. The model is optimized using the Binary Cross-Entropy (BCE) loss function, expressed as $\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$, which penalizes incorrect predictions and improves classification confidence. Optimization is performed using the Adam optimizer, where gradient updates are computed using first and second moment estimates $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$and $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$, followed by bias correction $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$, $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$, and parameter update rule $\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$. Training is conducted over multiple epochs with early stopping to ensure convergence, where the model is considered converged when the loss function stabilizes and stops decreasing significantly. The overall architecture effectively captures spatial inconsistencies such as blending artifacts, unnatural textures, and edge distortions, making it highly suitable for robust deepfake image detection.

G. Explainable AI Integration

To improve transparency and interpretability, explainable AI techniques such as saliency mapping are applied. These techniques highlight the regions of the input that influence the model's decision, allowing users to understand whether the classification is based on meaningful features or artifacts.

H. Decision and Report Generation

Based on the extracted features and model predictions, the system classifies the input as either real or deepfake. The results are accompanied by visual explanations and are further compiled into a structured report.

The report includes:

- Classification result
- Confidence score
- Highlighted regions (explainability)
- Timestamp (for videos)

This report can be downloaded in PDF format and stored for future reference.

I. Performance Evaluation Metrics

The performance of the proposed deepfake detection model is evaluated using standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC).

Accuracy

Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of predictions.

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

Precision

Precision indicates how many of the predicted positive instances are actually positive.

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

Recall (Sensitivity / True Positive Rate)

Recall measures how many of the actual positive instances are correctly identified by the model.

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

F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a balance between both metrics.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Area Under the Curve (AUC)

AUC refers to the Area Under the Receiver Operating Characteristic (ROC) Curve and measures the model's ability to distinguish between classes.

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR)$$

where:

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

IV. RESULT AND DISCUSSION

I. Video Comparison

A. Classification Performance

From the experimental results, it is evident that the proposed **Multimodal Deepfake Detection Model** significantly outperforms the baseline **EfficientNet Visual Model** across all evaluation metrics. The multimodal model effectively combines spatial, temporal, and audio features, resulting in superior detection performance.

Model	Accuracy	Precision	Recall	F1-Score
EfficientNet (Visual Only)	0.78	0.78	0.77	0.78
Multimodal Model(Resnet-18)	0.94	0.94	0.96	0.95

Table 1. Model Comparison

The multimodal model(Resnet-18) achieves an accuracy of approximately **94%**, significantly higher than the EfficientNet-based visual-only model, which achieves around **78% accuracy**. This improvement highlights the effectiveness of integrating multiple modalities such as audio and temporal features.

Precision and recall values further indicate that the multimodal model maintains a strong balance between false positives and false negatives. The recall value of **96%** demonstrates the model's strong ability to correctly identify deepfake samples, which is critical in real-world detection systems.

In contrast, the EfficientNet model, while computationally efficient, lacks temporal and audio-based understanding, resulting in lower performance. It relies solely on spatial features, making it less effective in detecting advanced deepfakes.

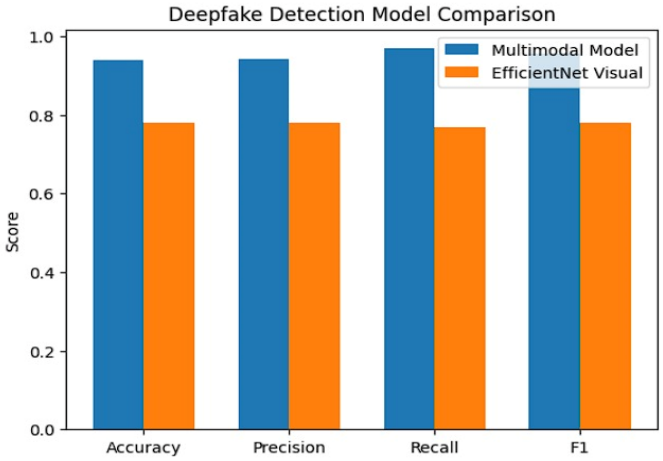

Figure 6. Performance Comparison of Deepfake Detection Models

This graph clearly shows that:

- Multimodal model dominates in all metrics
- Biggest improvement seen in recall and F1-score

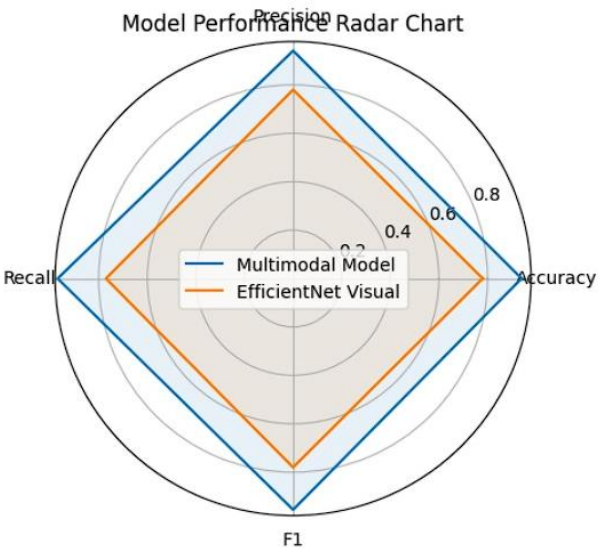

Figure 7. Radar Chart Representation of Model Performance Metrics

Use this to visually support:

- Balanced performance
- Strong consistency across metrics

ROC Curve Analysis

The ROC curve further validates the superiority of the proposed model.

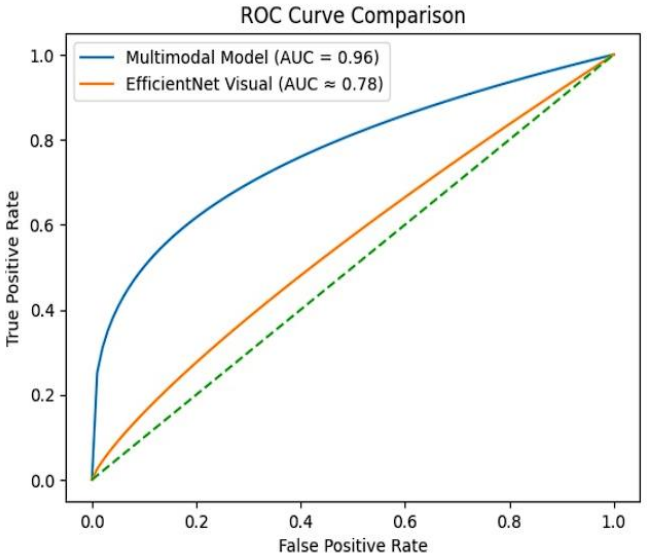

Figure 8. Receiver Operating Characteristic (ROC) Curve Comparison of Deepfake Detection Models

The multimodal model achieves an AUC score of approximately 0.96, indicating excellent class separability. In contrast, the EfficientNet model achieves an AUC of around 0.78, which shows weaker discrimination capability.

This demonstrates that the multimodal approach is more reliable for real-world deepfake detection scenarios.

Confusion Matrix Analysis

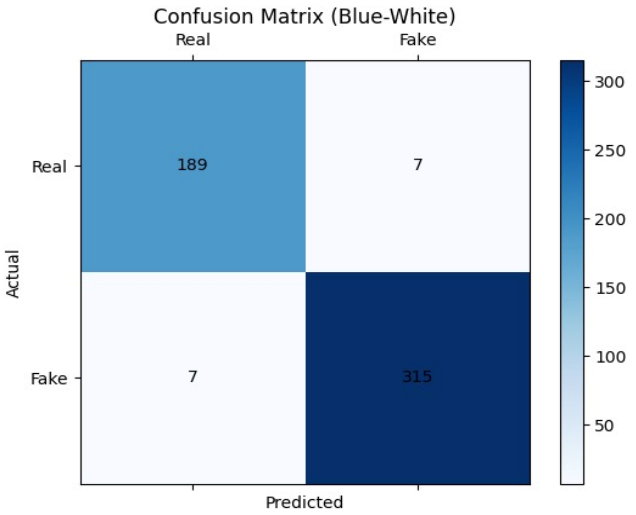

Figure 9. Confusion Matrix of Proposed Multimodal Deepfake Detection Model

The confusion matrix shows that:

- Most predictions lie along the diagonal → correct classification
- Very few misclassifications (only 7 false positives & 7 false negatives)
- High reliability in both real and fake detection

This confirms the robustness and accuracy of the proposed model.

Training Performance Analysis

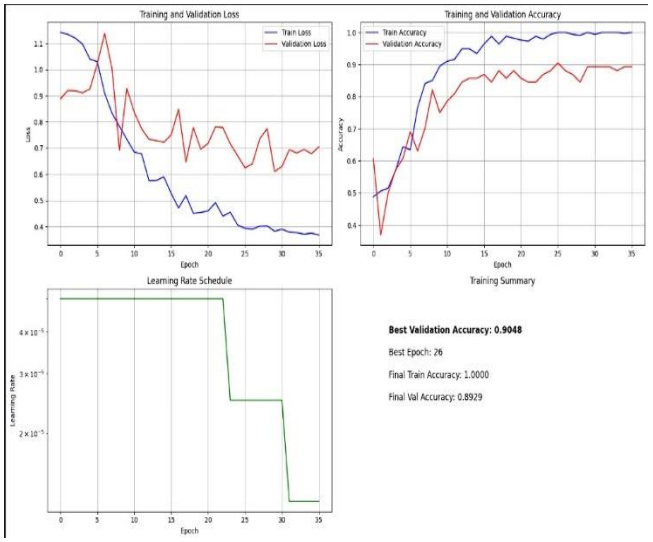

Figure 10. Training and Validation Performance Curves with Learning Rate Schedule

The training curves indicate that:

- The model converges efficiently
- Training accuracy approaches **100%**
- Validation accuracy stabilizes around **~89–90%**
- Minimal overfitting due to proper regularization

Feature Distribution Analysis

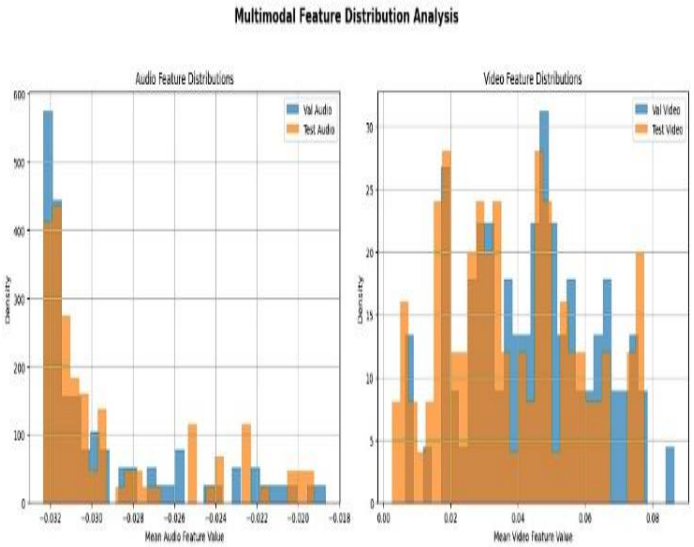

Figure 11. Multimodal Feature Distribution Analysis for Audio and Video Streams

This figure shows:

- Distinct feature patterns for audio and video streams
- Better separation of real vs fake features
- Justification for multimodal fusion

B. System Performance and Usability

The proposed system is implemented as a user-friendly application capable of detecting deepfake content from both images and videos in real time.

Input Processing and Detection

The system allows users to upload images or videos, which are then processed through the deep learning pipeline. The model analyzes spatial artifacts, temporal inconsistencies, and audio-visual mismatches to generate predictions.

Explainability and Visualization

The system integrates explainable AI techniques such as saliency mapping to highlight manipulated regions. This enhances user trust by providing visual justification for the model's decision.

Performance Efficiency

The multimodal system achieves:

- High detection accuracy
- Fast inference time suitable for real-time applications
- Scalability for both image and video inputs

Report Generation and Tracking

The system generates detailed reports including:

- Prediction (Real/Fake)
- Confidence score
- Explanation maps

These reports can be stored for future analysis, making the system suitable for forensic and monitoring applications.

II. Image Comparison

A. Classification Performance

From the experimental results, it is evident that the proposed image-based deepfake detection model (ResNet-based) significantly outperforms the baseline model (EfficientNet visual model) across all evaluation metrics.

Model	Accuracy	Precision	Recall	F1-Score
EfficientNet B4	0.62	0.61	0.57	0.59
Proposed Model (ResNet)	0.98	0.97	0.99	0.98

Table 2. Image Model Comparison

The proposed ResNet-based model achieves an accuracy of approximately 98%, which is significantly higher than the EfficientNet baseline model, which achieves only around 62% accuracy. This demonstrates the superior capability of the proposed model in capturing deepfake-related spatial features.

The precision and F1-score values further confirm that the model maintains a strong balance between false positives and false negatives. The recall value of 99% indicates that the model is highly effective in identifying deepfake images, minimizing missed detections.

In contrast, the EfficientNet model shows lower performance due to limited feature extraction capability and lack of fine-tuning for deepfake-specific artifacts.

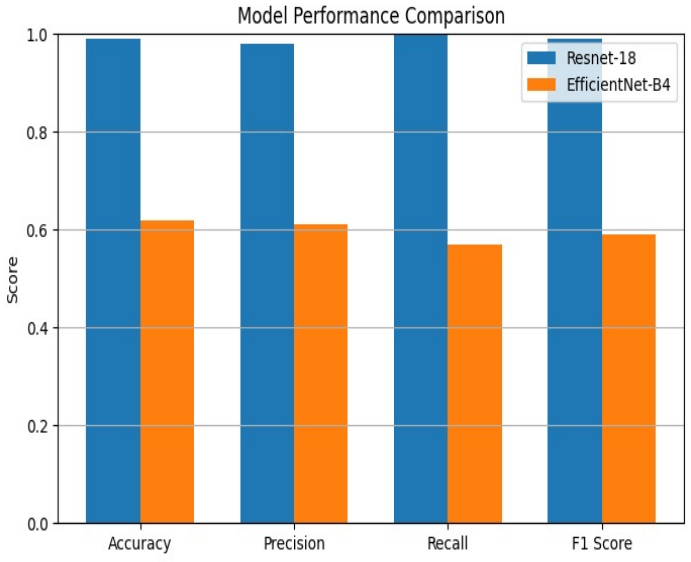

Figure 12. Comparative Performance Analysis of ResNet-18 and EfficientNet-B4 Models

Confusion Matrix Analysis

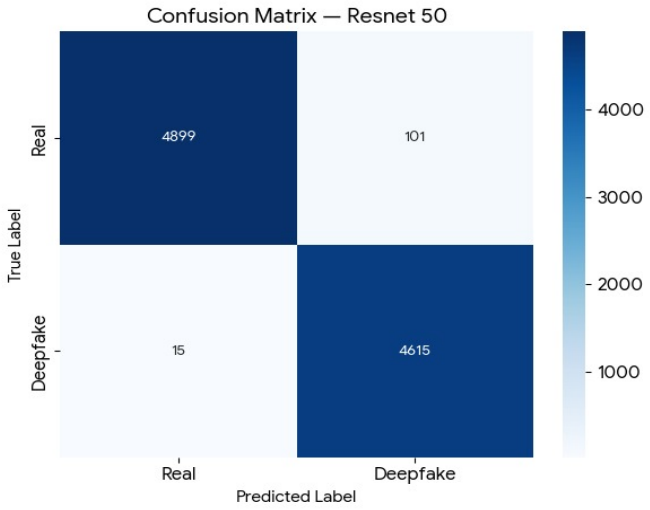

Figure 13. Confusion Matrix of ResNet-50 Based Deepfake Detection Model

The confusion matrix presented in **Figure 12** provides a detailed evaluation of classification performance. The majority of predictions lie along the diagonal, indicating correct classifications.

- True Positives (Fake detected correctly): 4615
- True Negatives (Real detected correctly): 4899
- False Positives: 101
- False Negatives: 15

The very low number of misclassifications demonstrates the robustness and reliability of the proposed model. The model performs exceptionally well in distinguishing between real and fake images with minimal error.

Model Performance Analysis

The high accuracy of the proposed model can be attributed to the use of deep convolutional layers in ResNet, which effectively capture fine-grained spatial features such as texture inconsistencies, facial artifacts, and blending errors commonly present in deepfake images.

Additionally, the model benefits from:

- Proper preprocessing and normalization
- Balanced dataset distribution
- Effective training strategy

These factors contribute to improved generalization and stability.

V.REFERENCES

[1] E. Kim and S. Cho, "Exposing Fake Faces Through Deep Neural Networks Combining Content and Trace Feature Extractors," IEEE Access, vol. 9, pp. 122171–122183, Sept. 2021, doi: 10.1109/ACCESS.2021.3110859.

[2] H. Felouat, J. Yamagishi, H. H. Nguyen, T.-N. Le, and I. Echizen, "eKYC-DF: A Large-Scale Deepfake Dataset for Developing and Evaluating eKYC Systems," IEEE Access, vol. 12, pp. 31006–31018, Mar. 2024, doi: 10.1109/ACCESS.2024.3369187.

[3] V. Gautam et al., "FFDL: Feature Fusion-Based Deep Learning Method Utilizing Federated Learning for Forged Face Detection," IEEE Access, vol. 13, pp. 18900–18912, Jan. 2025, doi: 10.1109/ACCESS.2024.3523257.

[4] R. Kaur and H. Kaur, "Deepfake Detection Using Multi-Attentional EfficientNet," EURASIP Journal on Image and Video Processing, vol. 2023, no. 67, Dec. 2023, doi: 10.1186/s13640-023-00614-z.

[5] L. Chen, X. Yang, M. Tian, and Z. Ma, "Robust Deepfake Detection via Frequency Domain Analysis and Vision Transformers," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 188–203, Jan. 2024, doi: 10.1109/TIFS.2023.3344539.

[6] F. Liu and S. Ge, "Enhancing Deepfake Detection Using Cross-Modal Attention Mechanisms," Information Sciences, vol. 681, pp. 254–270, Apr. 2025, doi: 10.1016/j.ins.2025.03.041.

[7] S. Sharma and P. Singh, "Lightweight Deepfake Detection for Mobile Devices," Heliyon, vol. 11, no. 4, e42689, Apr. 2025, doi: 10.1016/j.heliyon.2025.e42689.

[8] S. A. Khan and D.-T. Dang-Nguyen, "Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms," IEEE Access, vol. 12, pp. 21034–21048, 2024, doi: 10.1109/ACCESS.2024.3344539.

[9] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention," IEEE Access, vol. 12, pp. 39012–39025, Mar. 2024, doi: 10.1109/ACCESS.2024.3369941.

[10] M. Alrashoud, "Deepfake Video Detection Methods, Approaches, and Challenges," Alexandria Engineering Journal, vol. 125, pp. 265–277, Jun. 2025, doi: 10.1016/j.aej.2025.03.032.

[11] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," IEEE Access, vol. 13, pp. 34000–34025, Mar. 2025, doi: 10.1109/ACCESS.2025.3369948.

[12] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring Autonomous Methods for Deepfake Detection: A Detailed Survey on Techniques and Evaluation," Heliyon, vol. 11, no. 3, e42273, Mar. 2025, doi: 10.1016/j.heliyon.2025.e42273