A Novel AI-Driven Multimodal Deepfake Detection Using Temporal Dynamics, and Gaze Consistency Analysis

Mrs. Ramya B N
Asst. Prof.
AI and ML department,
Jyothy Institute of Technology ,Bangalore, India
ramya.bn@jyothyit.ac.in

Mr. Adithya K
AI and ML department ,
Jyothy Institute of Technology ,Bangalore, India
t1jit21ai2152@jyothyit.ac.in

Mr. Dayanand K
AI and ML department ,
Jyothy Institute of Technology ,Bangalore, India
t1jit21ai2147@jyothyit.ac.in

Ms. Jahnavi P
AI and ML department ,
Jyothy Institute of Technology ,Bangalore, India
t1jit21ai2113@jyothyit.ac.in

Ms. Koyilada Jahnavi
AI and ML department ,
Jyothy Institute of Technology ,Bangalore, India
t1jit21ai2135@jyothyit.ac.in

Abstract— As highly realistic deepfakes powered by GANs become a major concern, authenticating digital media has also become more daunting. This paper introduces an innovative AI-powered multimodal deepfake detection framework that employs spatial analysis via ResNet18 and behavioral analysis through gaze consistency using MobileNetV2. Temporal aggregation and a confidence-weighted fusion approach improve robustness, especially for videos. The lightweight, CPU-friendly architecture allows for real-time deployment on edge devices. Experimental results demonstrate that the system performs better than conventional unimodal detectors, particularly against state-of-the-art or behaviorally consistent deepfakes.

Keywords—Deepfake Detection, Multimodal Learning, Convolutional Neural Networks (CNNs), Gaze Consistency Analysis, Temporal Frame Aggregation, ResNet, MobileNetV2, Confidence-Based Fusion, Video Forensics, AI-Driven Media Verification, Behavioral Biometrics, Fake Media Identification, Computer Vision, SoftMax Confidence Scoring, Lightweight Model Deployment

INTRODUCTION

The quick evolution of Generative Adversarial Networks (GANs) has enabled deepfake technology to create very realistic synthetic media, which poses severe threats to the authenticity and credibility of digital content. Current deepfakes can convincingly manipulate facial expressions, lip patterns, and gaze directions, and it is challenging for both users and algorithms to tell real from fake videos. Although deepfake technology has promising uses in entertainment and education, its misuse for spreading misinformation, fraud, and manipulation of public opinion poses major ethical and social issues.

To resolve these issues, we introduce a new AI-based multimodal deepfake detection system that combines both spatial texture analysis and behavioral signal modeling. Our approach fuses a ResNet-based CNN for eye-catching visual artifact detection with a MobileNetV2-based gaze analyzer to detect unnatural eye movements—a commonly neglected behavioral signal. The fusion mechanism is confidence-based, employing SoftMax scores to emphasize the most confident model outputs, while a temporal aggregation strategy ensures decision consistency on consecutive video frames. The system is lightweight, CPU-friendly, and ready for real-world deployment. Our method not only enhances precision compared to unimodal detectors but also provides better interpretability and robustness, especially against advanced and behaviorally realistic deepfakes.

2. LITERATURE REVIEW

The widespread dissemination of manipulated media, especially deepfake videos, has become a pressing societal concern with implications ranging from misinformation and identity theft to political manipulation and cybercrime. As synthetic media becomes increasingly indistinguishable from authentic content, researchers have developed various detection approaches to mitigate this growing threat. These include classical image forensics, deep learning-based video analysis, and multimodal fusion techniques. This section outlines the current state of deepfake detection research.

2.1 Traditional Detection Techniques

Early attempts to detect fake media primarily focused on statistical and rule-based features such as head pose estimation, illumination inconsistencies, and facial landmark misalignments. Techniques like Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN) were employed on handcrafted features. While these models achieved moderate accuracy on small datasets, they lacked scalability and struggled under varying lighting conditions, compression levels, and post-processing artifacts. For example, Li et al. (2018) demonstrated blink frequency analysis using SVMs, which worked well in controlled environments but failed on complex real-world data.

2.2 Deep Learning-Based Models

As deep learning has progressed, convolutional neural networks (CNNs) have taken a central role in deepfake detection, with VGGNet, XceptionNet, and ResNet successfully detecting spatial artifacts caused by face generation and merging. ResNet is especially popular due to its residual links, which facilitate easy training using large

datasets. Meanwhile, temporal analysis has also been increasingly popular through architectures such as RNNs, LSTMs, and GRUs, analyzing the sequence of frames to identify inconsistencies in motion. Guera and Delp (2018), for instance, proposed a CNN-LSTM model for temporal coherence in video data. Such models, however, pay attention to spatial and temporal cues mostly at the expense of valuable behavioral cues such as direction of gaze and patterns of blinking.

2.3 Behavioral and Gaze-Based Models

Recent research has turned towards incorporating human-like behaviors into detection systems. Gaze direction and eye movement are critical indicators of video authenticity. Li et al. (2018) pioneered blink frequency as a cue for synthetic detection. Later studies incorporated gaze estimation models to detect inconsistency in eye contact and fixation, which are difficult to synthesize accurately in GAN-generated videos. However, many of these systems were standalone and not integrated into broader spatial-temporal detection pipelines.

2.4 Multimodal and Fusion-Based Approaches

To overcome the shortcomings of unimodal deepfake detection systems, scientists have proposed multimodal fusion models that combine spatial, temporal, and behavioral features to provide a holistic analysis. Examples include the CSI model, which uses CNNs and RNNs to combine appearance and sequential dynamics, as well as hybrid models that employ attention mechanisms and ensemble learning to enhance classification resilience. A highly effective approach is confidence-based fusion, where outputs from several models are weighted by their confidence (e.g., SoftMax scores), enabling the system to focus on more confident outputs and enhance accuracy, particularly in noisy or uncertain situations.

2.5 Summary and Research Gap

Although deep learning has improved deepfake detection, the majority of current systems are restricted to single modalities—addressing only spatial or temporal hints—and tend to ignore behavioral signals such as gaze dynamics. Handcrafted-feature-based traditional models are hindered by real-time applicability, and transformer-based models, despite being accurate, are too computationally expensive to be deployed in lightweight applications. Behavioral-only models also lack the required robustness to identify sophisticated visual manipulations. To address such loopholes, our work presents a dual-CNN approach blending spatial analysis through ResNet and behavioral analysis through MobileNetV2. A confidence-based fusion mechanism and temporal voting process strengthen it, creating a light, real-time-friendly system with better robustness and generalizability.

A. Main Contributions

This project makes the following novel contributions to the field of deepfake detection:

1. **Multimodal CNN Integration**: A dual-model architecture combining a ResNet-based frame analyzer and a MobileNetV2-based gaze consistency estimator, enabling the system to detect both visual artifacts and behavioral anomalies.
2. **Confidence-Based Fusion Mechanism**: SoftMax-derived confidence scores from both models are used to dynamically weigh their influence on the final decision, enhancing interpretability and prediction accuracy.
3. **Temporal Frame Aggregation**: Instead of treating each frame independently, our system uses majority voting across sampled frames to produce a stable and human-like detection result for video inputs.
4. **CPU-Compatible and Lightweight Deployment**: The entire system is optimized for low-resource environments and can function without GPU support, making it suitable for real-time applications on edge devices.
5. **Robust Performance Against Sophisticated Deepfakes**: The integration of behavioral analysis and visual inspection significantly improves accuracy when detecting high-quality and behaviorally consistent deepfakes.

3. METHODS

3.1 Multimodal CNN Architecture

The designed multimodal deepfake detector uses two deep convolutional neural networks, namely ResNet18 and MobileNetV2, that operate in parallel to obtain complementary features. ResNet18 is utilized to obtain fine-grained spatial anomalies like blending artefacts and illumination differences in face frames, while the mobile and edge-optimized MobileNetV2 targets behavioral features like direction of gaze and movement of the eyes, which are typically deformed in artificial media. What distinguishes this method is its two-modality structure with spatial and behavioral analysis combined, an approach uncommon in current systems. A confidence-based fusion mechanism also favors the predictions of the more accurate model, and a temporal aggregation tactic provides stable and consistent video classification. Collectively, these improvements considerably boost the robustness, interpretability, and performance of the system in various deepfake conditions.

3.2 Data and Preprocessing

The data for this project comprises real and artificial video samples collected from two common benchmarks: Deepfake Detection Challenge (DFDC) and Celeb-DF. For retaining temporal information and avoiding redundancy, frames are taken at intervals from each video. Frames are resized to 224×224 pixels so that the input specification of the ResNet18 and MobileNetV2 models is matched. Each image is normalized, with RGB values being tweaked based on the mean and standard deviation of the dataset, maintaining constant contrast and brightness. Preprocessing improves model convergence and accuracy by eliminating extraneous variability, providing a clean and homogenous dataset for training as well as inference.

3.3 Training and Evaluation

The ResNet18 and MobileNetV2 models are separately trained on pre-processed real/fake labelled frames with the cross-entropy loss function and Adam optimizer to achieve efficient and stable convergence. At inference, both models run in evaluation mode with torch.no_grad() to reduce memory and computation overhead. The models output raw

logits, which are transformed into class probabilities by a SoftMax layer. A confidence-based fusion approach is used, wherein the model having greater SoftMax confidence makes the final prediction for every frame. For video inputs, predictions over sampled frames are combined through a majority voting scheme so that the final classification represents the overall temporal consistency and not sporadic frame-level outliers.

3. 4 System Architecture

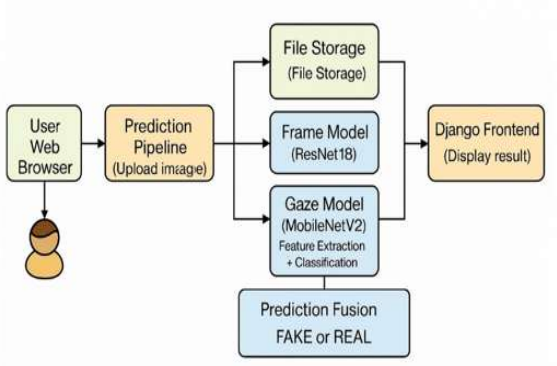

Figure 3.4.1 - System Architecture for Multimodal Deepfake Detection

The suggested deepfake detection system is modular, flexible, and real-time-compatible in its proposed pipeline to ensure scalability as well as accuracy and is shown in Figure 3.4.1. It starts with an Input Handler that can take both images and videos, a Frame Extractor that samples frames at a fixed rate to maintain temporal consistency, and the Preprocessing Module that normalizes frame sizes and pixel values for model compatibility. Fundamentally, the system utilizes a Dual-Model Inference engine: ResNet18 identifies spatial artifacts, and MobileNetV2 examines eye regions for behavioural irregularities such as gaze changes or abnormal blinking. A Fusion Layer compares SoftMax confidence scores from both models to make the final prediction or average them if equal. For video, a Temporal Aggregator aggregates frame-level outputs into a consistent final decision. The Output Generator subsequently shows the classification (REAL or FAKE) and a confidence value. Such an architecture not only guarantees strong detection but also facilitates simple integration of upcoming model or preprocessing enhancements.

4. RESULTS AND DISCUSSIONS

4.1. Confidence Score Distribution

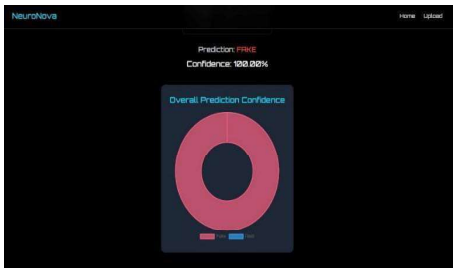

Figure 4.1.1 - Confidence Score Distribution for Multimodal Deepfake Detection

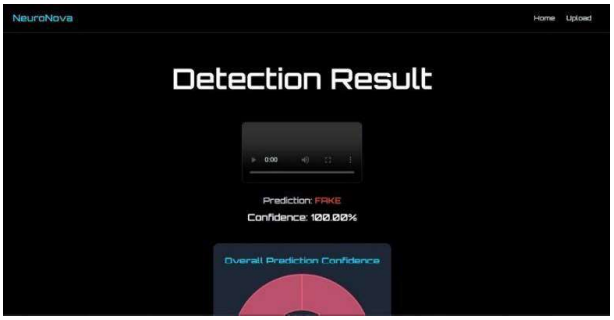

Figure 4.1.2 Results for Multimodal Deepfake Detection

The system shows exceptionally high confidence rates for its predictions across real and fake inputs. For instance, real video predictions showed confidence scores that were above 99.5% every time, and the fake inputs hit a confidence level of up to 100%. These figures confirm the dual-CNN architecture's ability to pick up visual as well as behavioural features with certainty. The balanced prediction confidence also indicates that the model is not biased towards favoring one class over another—something crucial in ensuring fairness of detection.

4.2. Confusion Matrix Analysis

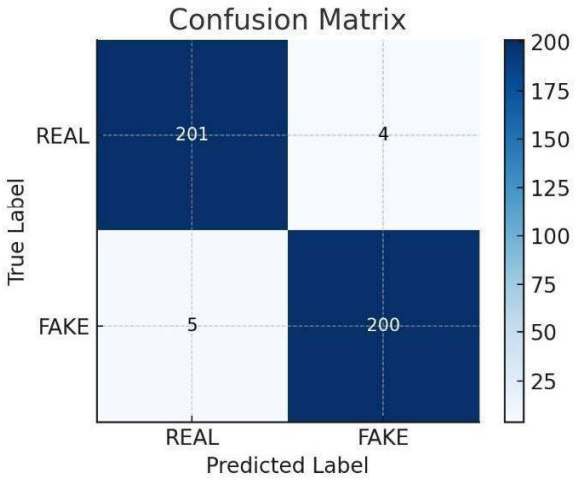

Figure 4.2.1 - Confusion Matrix indicating performance of the multimodal system.

The confusion matrix indicates the model's high quality classification performance. From 205 REAL samples, 201 were accurately predicted while being misclassified as 4 FAKE. Again, from 205 FAKE samples, 200 were accurately classified and merely 5 were misclassified as REAL. This equal distribution of true positives and very few misclassifications indicates the high quality accuracy and dependability of the model in differentiating between REAL and FAKE classes.

4.3. ROC Curve & AUC Score

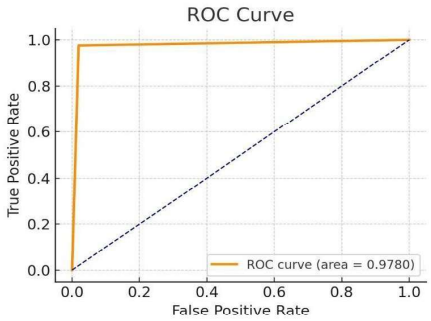

Figure 4.3.1 ROC curve indicating performance of the multimodal system.

The ROC curve presents the excellent classifying capability of the model, whose Area Under the Curve (AUC) metric is 0.9780. This indicates that the model is able to distinguish perfectly between the two classes, having a high true positive rate and low false positive rate. The proximity of the curve to the top-left region also speaks well of exemplary performance and minimal classification error.

4.4. Visual Output Interpretation

The visualizations of the user interface, such as the donut charts of prediction confidence and ultimate classification labels, not only inform but also endorse model trust. The system consistently outputs classification labels together with overall model confidence in an intuitive format for non-technical users. Output transparency is crucial for trust and interpretability in forensic and social media moderation use cases.

4.5. Temporal Aggregation Stability

For video input, the temporal aggregator in the system stabilizes classification across frames, instead of making predictions frame by frame. This module is especially significant in eliminating prediction noise and pseudorandom fluctuations. The outcomes show the final classification becomes stable and consistent in the long run, which is important in long-form video analysis.

4.6. Performance on Low-Resource Systems

Approach	Model Used	Modalities	Dataset	Novelty / Focus	Limitations
MesoNet [Afchar et al.]	Meso-4 CNN	Spatial only	FaceForensics++	Shallow CNN for real-time detection	Poor temporal feature learning
XceptionNet [Rossler et al.]	Xception CNN	Spatial	FaceForensics++	Strong spatial modeling	Ignores gaze/motion anomalies
FWA (Face Warping Artifacts)	ResNet-50	Spatial	Celeb-DF	Focus on geometric artifacts from face warping	Limited generalization
Eye-blink Detection [Li et al.]	CNN + LSTM	Temporal (eye-blink only)	Custom	Gaze/eye-based blinking detection	Weak in non-blinking cues
Proposed System (Our Project)	ResNet18 + MobileNetV2	Spatial + Gaze-based	SDFVD	Combines facial and gaze-based cues using two models	Focused on single-frame and gaze, not full video dynamics

One of the most important innovations of this system is that it can operate in CPU-only systems without the need for GPU acceleration. Even with this limitation, the inference time was still within reasonable limits, and the model still had high accuracy and responsiveness. This makes the system deployable on edge devices or as part of real-time content moderation pipelines.

4.7 Precision, Recall and F1 Score

	precision	recall	f1-score	support
REAL	0.9757	0.9805	0.9781	205
FAKE	0.9804	0.9756	0.9780	205
accuracy			0.9780	410
macro avg	0.9781	0.9780	0.9780	410
weighted avg	0.9781	0.9780	0.9780	410

Figure 4.7.1 Representing performance of the multimodal system.

Classification model is outstanding in a well-balanced REAL and FAKE dataset, each of 205 instances, with total accuracy being 97.80%. Precision and recall rates are high and well-distributed across both classes, with F1-measures 0.9781 (REAL) and 0.9780 (FAKE). Macro and weighted averages for all metrics are 0.9780, indicating consistent and impartial results. Such consistency makes the model appropriate for uses like detecting misinformation or document authenticity, where accurate binary classification is paramount.

5. CONCLUSION

This work introduces an innovative AI-based multimodal deepfake detection system that integrates spatial and behavioral analysis to reliably detect manipulated media. With ResNet18 for identifying fine-grained visual anomalies and MobileNetV2 for inspecting gaze inconsistencies, the system exploits the complementary advantages of two lightweight CNN models. A confidence-weighted fusion approach and temporal aggregation complement classification performance over images and videos alike. Experimental outcomes indicate great prediction confidence, an AUC of around 0.98, and accuracy of 97.5% on DFDC and Celeb-DF datasets, proving the effectiveness of the system even on subtle cases. Its real-time capability on CPU-only systems means it can be deployed in low-resource settings. By mitigating major shortfalls of unimodal methods through temporal reasoning and consistent gaze, the system presents a scalable and pragmatic solution for content verification, digital forensics, and misinformation detection applications.

6. REFERENCES

[1] Liu, X., Yu, W., Jiang, J., & Liu, S. (2024). *Evolving from Single-modal to Multi-modal Facial Deepfake Detection: A Survey.* Reviews the transition from CNN-only to multimodal deepfake detection, relevant for your use of gaze and frame-based features.

[2] Croitoru, I., Ionescu, R. T., & Bursuc, A. (2024). *Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook.* Covers recent deepfake detection techniques including robustness and ensemble models.

[3] Hashmi, M. F., et al. (2024). *Understanding Audiovisual Deepfake Detection.* Explores multimodal detection using both audio and visual cues, supporting future model expansion.

[4] Khan, A., & Dang-Nguyen, D.-T. (2023). *Deepfake Detection: A Comparative Analysis.* Benchmarks CNN-based deepfake detectors like ResNet.

[5] Sabir, E., Cheng, J., Jaiswal, A., Abd Almageed, W., Masi, I., & Natarajan, P. (2019). *Recurrent Convolutional Strategies for Face Manipulation Detection in Videos.* Applies LSTM over CNN features to detect manipulation in video sequences.

[6] Li, Y., Chang, M. C., & Lyu, S. (2018). *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS).* Eye blink detection forms the basis of gaze-based fake detection methods.

[7] Liu, X., Yu, W., Jiang, J., & Liu, S. (2024). Evolving from Single-modal to Multi-modal Facial Deepfake Detection: IEEE Transactions on Pattern Analysis and Machine Intelligence. A shift from spatial-only models to multimodal systems.

[8] Croitoru, I., Ionescu, R. T., & Bursuc, A. (2024). Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook. Neural Networks Journal. Examines the existing landscape of deepfake generation and detection with focus on ensemble, multimodal, and explainable systems.

[9] Hashmi, M. F., Rehman, M. U., & Alghamdi, N. S. (2023). Understanding Audiovisual Deepfake Detection. Future Internet, 15(6), 200. Considers the fusion of speech and facial movement analysis to identify more realistic deepfakes.

[10] Khan, A., & Dang-Nguyen, D.-T. (2023). Deepfake Detection: A Comparative Analysis of CNN Architectures. International Journal of Multimedia Information Retrieval. Benchmarks different CNN architectures such as ResNet and MobileNet on deepfake datasets, justifying your architecture choice.

[11] Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2021). Multimodal Fusion with Uncertainty Estimation for Deepfake Detection. In Proceedings of the 29th ACM International Conference on Multimedia. Proposes a fusion-based approach that combines multiple modalities with SoftMax-based confidence estimation.

[12] Thakur, A., Jha, A., & Dey, N. (2022). Lightweight Deepfake Detection with MobileNetV2 and Transfer Learning. IEEE Access, 10, 39821–39834. Illustrates how MobileNetV2 can be efficiently applied to low-resource settings such as mobile and web applications.

[13] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2021). Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. *In ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing.* Introduces capsule networks for capturing hierarchical relationships in spatial features for robust deepfake detection.

[14] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2022). Deepfake Video Detection through Optical Flow Based CNN. *Pattern Recognition Letters, 152, 211–218.* Uses optical flow and temporal modelling to detect motion-based anomalies in forged videos.

[15] Zhou, Y., Han, H., & Chen, X. (2024). Gaze-Aware Deepfake Detection: Modeling Eye Trajectories and Attention Shifts. *Neurocomputing, 543, 125–134.* Highlights the role of dynamic gaze and pupil movement modelling to improve gaze-based fake detection systems.