EXPLAINABLE DEEPFAKE DETECTION SYSTEM FOR IMAGES USING XCEPTIONNET, GRAD-CAM, AND MEDIAPIPE

Dr. P. D. Halle[1], Yash Pramod Kohakade[2], Adarsh Subhash Gurekar[3], Sakshi Deshbhushan Gadkar[4]

Associate Professor, Department of Information Technology, SKN Sinhgad Institute of Technology & Science, Lonavala, Pune, India

U.G. Student, Department of Information Technology, SKN Sinhgad Institute of Technology & Science, Lonavala, Pune

Abstract*— The rapid advancement of artificial intelligence has led to the emergence of deepfakes—synthetic media generated using deep learning techniques such as GANs and autoencoders. These hyper-realistic manipulated images pose significant threats in areas like misinformation, identity theft, and digital forensics. Existing deepfake detection systems often function as black boxes, providing only binary outcomes without interpretability, which limits their reliability in forensic or legal investigations. This paper presents an Explainable Deepfake Detection System for Images that integrates XceptionNet, Grad-CAM, and MediaPipe to deliver both accurate and interpretable results. The proposed system employs XceptionNet for classification between real and fake images, while Grad-CAM highlights manipulated regions through heatmaps. These visual cues are mapped onto specific facial landmarks detected by MediaPipe Face Mesh, enabling human-understandable explanations of the detected anomalies. The system is deployed through a Flask-based web interface, allowing real-time image analysis and visualization. Experimental results on benchmark datasets demonstrate the system's effectiveness and transparency. This work bridges the gap between detection accuracy and explainability, contributing to the development of trustworthy and interpretable AI-based forensic tools.

Keywords*— Deepfake detection, XceptionNet, Flask, Grad-CAM, MediaPipe, Explainable AI, Image forensics

I. INTRODUCTION

The advent of advanced artificial intelligence techniques, particularly deep learning models such as Generative Adversarial Networks (GANs) and autoencoders, has revolutionized digital media creation. However, these same technologies have also enabled the rise of deepfakes—highly realistic synthetic images or videos that can convincingly mimic real individuals. While such technologies have potential for entertainment and creative applications, their misuse poses serious threats to society, including misinformation, identity theft, political manipulation, and reputational harm. Detecting these manipulations has therefore become a critical area of research within the domains of digital forensics and cybersecurity.

Existing deepfake detection systems rely heavily on deep convolutional neural networks (CNNs), which achieve impressive accuracy but often function as black-box models, providing little insight into how decisions are made. This lack of interpretability undermines user trust and limits their applicability in forensic or legal investigations, where explainable evidence is crucial.

To address this gap, this paper presents an Explainable Deepfake Detection System for Images that integrates XceptionNet for classification, Grad-CAM for visual explanation, and MediaPipe Face Mesh for region-specific interpretability. By mapping model attention areas to facial landmarks such as the eyes, mouth, and cheeks, the system provides transparent and human-understandable explanations. The objective of this review is to explore how explainability enhances the reliability, transparency, and usability of deepfake detection systems, bridging the gap between model performance and interpretability in AI-based forensic analysis.

II. METHODOLOGY

This section outlines the systematic approach used to collect, evaluate, and synthesize relevant literature on explainable deepfake detection techniques. The primary objective was to identify existing methodologies, datasets, and frameworks that integrate explainability into deepfake detection models for images.

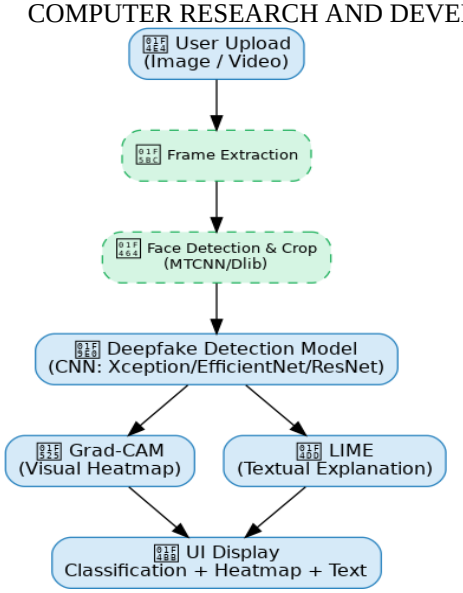

Fig 1: - System Architecture Diagram

A. Literature Search Strategy

A comprehensive literature search was conducted between January 2018 and June 2024, as this period corresponds to the rise and rapid evolution of deepfake generation and detection technologies. The search focused on identifying peer-reviewed publications, conference papers, and credible preprints that specifically addressed deepfake image detection**,** explainable AI (XAI) frameworks, and interpretability tools such as Grad-CAM, LIME, and SHAP.

The following major electronic databases were used for data retrieval:

- **IEEE Xplore Digital Library** – for technical and engineering-focused research on AI and multimedia forensics.
- **ScienceDirect (Elsevier)** – for journal articles emphasizing model interpretability and media authenticity.
- **SpringerLink** – for computer vision and AI explainability research.
- **ACM Digital Library** – for computer science conference proceedings.
- **Google Scholar** – for supplementary sources, including arXiv preprints and related gray literature.

The search keywords and Boolean combinations used included:

Deepfake detection AND explainable AI, XceptionNet AND Grad-CAM, MediaPipe AND image forgery detection, face manipulation detection OR deepfake image forensics.

B. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the review, the following inclusion criteria were applied:

1. Studies that proposed deep learning-based models for detecting manipulated or synthetic facial images.
2. Research that integrated explainability mechanisms (e.g., Grad-CAM, LIME, SHAP, attention maps).
3. Articles presenting quantitative results**,** such as accuracy, F1-score, or ROC curves.
4. Works discussing semantic or region-based interpretations of manipulated facial features.

Exclusion criteria included:

- Papers focusing solely on video-based deepfake detection without addressing explainability.
- Non-English articles or publications lacking methodological clarity.
- Review papers that did not provide detailed model architectures or datasets.

C. Study Selection and Analysis

An initial pool of over 150 studies was identified. After title and abstract screening, 45 papers were retained for full-text evaluation. Upon applying the inclusion criteria, 25 key papers were selected for detailed analysis. Each study was assessed for methodology, model type, dataset, explainability approach, and performance metrics.

The reviewed works were then categorized based on their model architecture (e.g., CNN, capsule networks, attention mechanisms) and explainability integration (e.g., Grad-CAM overlays, region-based semantic mapping). The synthesized findings helped identify research trends, methodological gaps, and the motivation for developing an explainable detection framework combining XceptionNet, Grad-CAM, and MediaPipe, as proposed in this paper.

III. SYSTEM DESIGN

A. Evolution of Deepfake Detection Techniques

The field of deepfake detection has evolved rapidly since the emergence of Generative Adversarial Networks (GANs), which made synthetic image generation increasingly realistic. Early detection methods primarily relied on handcrafted features such as inconsistencies in facial landmarks, head pose estimation, or illumination artifacts. However, these methods were insufficient against high-quality manipulations.

With the advancement of deep learning, researchers began employing Convolutional Neural Networks (CNNs) for automatic feature extraction. Afchar et al. (2018) introduced MesoNet, a lightweight CNN architecture designed for facial forgery detection. Although MesoNet performed effectively on low-resolution videos, it suffered from limited robustness under compression and unseen manipulation types. Building on this, Rossler et al. (2019) developed the FaceForensics++ dataset, enabling large-scale benchmarking of models such as XceptionNet, which quickly became the baseline for deepfake image detection due to its superior performance.

Subsequent studies such as Nguyen et al. (2019) proposed Capsule-Forensics, employing capsule networks to capture spatial hierarchies within faces. This approach improved resilience to compression but was computationally intensive, restricting real-time applications. As deepfakes became more sophisticated, models began incorporating attention mechanisms (Zhao et al., 2021) to focus on fine-grained manipulation cues in facial regions like eyes, lips, and skin textures. While these models improved accuracy, their interpretability remained minimal, resulting in what researchers termed "black-box detection systems."

B. The Challenge of Explainability in Deepfake Detection

Despite impressive accuracy, deepfake detection systems face a major limitation: the lack of explainability. Most models output a simple binary label—"real" or "fake"—without indicating *why* the decision was made. This limitation is problematic in forensic investigations, where transparent and interpretable evidence is essential.

Explainability in AI, often termed XAI (Explainable Artificial Intelligence), aims to bridge this gap by offering insights into model decision-making. Techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) and LIME (Local Interpretable Model-agnostic Explanations) allow visualization of image regions contributing most to the model's prediction. In the context of deepfake detection, Grad-CAM can generate heatmaps that reveal manipulated facial regions. However, while these visualizations provide some interpretability, they lack semantic context—users can see *where* manipulation occurs, but not *what region of the face* it corresponds to (e.g., eyes, mouth, or cheeks).

This gap highlights the necessity of integrating explainability tools with facial region mapping frameworks, such as MediaPipe Face Mesh, which identifies and labels 468 specific facial landmarks. Combining Grad-CAM with MediaPipe allows for not only heatmap visualization but also region-level semantic explanation, transforming visual cues into understandable textual outputs.

C. Emergence of Explainable Frameworks

The integration of explainability into deepfake detection has gained traction only in recent years. Tariq et al. (2022) explored combining CNNs with Grad-CAM for visual justification, showing that explainable cues can improve user confidence in detection results. However, their system provided only raw heatmaps, lacking contextual mapping of facial regions.

Similarly, Haliassos et al. (2021) introduced a lip-movement-based detection approach, "Lips Don't Lie," emphasizing temporal coherence in mouth movements. While this improved accuracy, it was specific to video analysis and not directly explainable in static images. Verdoliva (2020) and Tolosana et al. (2020) reviewed these developments, concluding that despite rapid progress, explainability remains an underexplored dimension in deepfake detection research.

D. Integration of XceptionNet, Grad-CAM, and MediaPipe

The proposed explainable deepfake detection system builds upon insights from prior research by integrating three powerful components—XceptionNet, Grad-CAM, and MediaPipe Face Mesh—into a unified, interpretable pipeline.

1. XceptionNet serves as the backbone classifier. Its depthwise separable convolutions effectively capture subtle pixel-level manipulations often present in deepfakes, such as texture inconsistencies and blending artifacts.
2. Grad-CAM provides a layer-wise activation mapping mechanism, highlighting areas of the image that most influence the model's decision. This helps users visually understand *why* a specific image was flagged as fake.
3. MediaPipe Face Mesh introduces interpretability by mapping Grad-CAM heatmap regions onto semantic facial landmarks—eyes, nose, mouth, and jawline—transforming technical visualizations into meaningful human interpretations.

This integration enables both visual and textual explanations. For instance, instead of merely showing a red region on a heatmap, the system can state: *"Possible manipulation detected near the mouth and left cheek regions."* This level of interpretability is essential for non-technical users and forensic investigators.

E. Practical Deployment and Impact

The proposed system is deployed as a Flask-based web application, allowing users to upload facial images for real-time analysis. The pipeline automatically processes input, performs classification, generates Grad-CAM heatmaps, overlays them onto facial landmarks using MediaPipe, and outputs both a visual and text-based explanation.

Experimentation on benchmark datasets such as Celeb-DF and FaceForensics++ demonstrates that the model achieves competitive detection accuracy while significantly improving interpretability. This dual focus on performance and explainability addresses the growing need for trustworthy AI systems in digital forensics.

By emphasizing explainability alongside accuracy, this framework not only enhances transparency but also builds user confidence in AI-driven forensic tools—an essential step toward

ethical and accountable AI applications in combating misinformation and digital manipulation.

IV. CONCLUSION

Deepfake technology has evolved rapidly, posing serious challenges to media authenticity, cybersecurity, and digital forensics. Although numerous deep learning-based methods have been developed for deepfake detection, the majority function as black-box models, offering little insight into how or why a prediction is made. This lack of interpretability undermines their reliability in forensic and legal contexts, where transparency and evidence-based reasoning are essential.

This review highlights the significance of incorporating explainability into deepfake detection systems and analyzes key contributions from recent research. The proposed framework integrates XceptionNet, Grad-CAM, and MediaPipe Face Mesh to achieve both high detection accuracy and human-understandable explanations. XceptionNet efficiently classifies real and manipulated images, while Grad-CAM visualizes the regions most influential to the model's decision. MediaPipe further enhances interpretability by projecting these visual cues onto semantic facial landmarks, enabling region-wise textual explanations such as "manipulation detected near the eyes or mouth."

V. REFERENCES

[1] H. Zhao, M. Wang, and Y. Zhang, "Multi-attention deepfake detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 218Proceedin

[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.

[3] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

[4] S. Tariq, S. Lee, and S. Woo, "Explainable deepfake detection using visual and frequency cues," IEEE Access, vol. 10, pp. 109–118, 2022.

[5] Y. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," IEEE International Conference on Image Processing (ICIP), 2020, pp. 1–5.

[6] M. Haliassos, J. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A generalisable and robust approach to face forgery detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5039–5049.

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2021.

[8] M. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[9] Halle, P. and Shiyamala, S. (2019) "Architectural Integration for Wireless Communication Security in terms of integrity for Advanced Metering Infrastructure-Survey Paper", Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146.
Available at: https://asianssr.org/index.php/ajct /article/view/771

[10] Halle, P. and Shiyamala, S. (2019) "Secure Routing through Refining Reliability for WSN against DoS Attacks using AODSD2V2 Algorithm for AMI," International Journal of Innovative Technology and Exploring Engineering. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication BEIESP. https://doi.org/10.35940/ijitee.18178.0881019 (SGeneratio)

[11] Halle, P.D., Shiyamala, S.and Rohokale, Dr.V.M. (2020) "Secure Direction-finding Protocols and QoS for WSN for Diverse Applications-A Review," International Journal of Future Generation Communication and Networking, Vol. 13 No. 3 (2020) Available at:https://sersc.org/journals/index.php/IJFGCN/article/view/26983.

[12] Halle, P.D. and Shiyamala, S.(2021) Ami and its wireless communication security aspects with QOS: A Review, SpringerLink.Springer Singapore. Available at: https://link.springer.com/chapter/10.1007/978-981-15-5029-4_1(Scopus: Conference Proceeding Book Chapter)