International Conference on Advances in Modern Technology of Research in Engineering Field (AIMTREF) April, 2026
Organized By: SKN Sinhgad Institute of Technology& Science, Kusgaon (Bk), Lonavala, Pune

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Explainable Deepfake Detection System Using Xceptionnet, Grad-Cam, and Mediapipe

Mr.S.P.Gunjal, Yash Kohakade, Adarsh Gurekar, Sakshi Gadkar
Department of Information Technology, SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra

Abstract- Deepfake technology has emerged as a major threat to digital media authenticity in recent years. Deepfakes are manipulated images or videos generated using deep learning models that can closely resemble real human faces. These fake media contents are often misused for spreading misinformation, identity theft, and reputation damage. Although many deep learning models have been developed for deepfake detection, most of them work as black-box systems and do not provide any explanation for their predictions. This paper presents an Explainable Deepfake Detection System for facial images using XceptionNet, Grad-CAM, and MediaPipe Face Mesh. The proposed system is implemented as a Python Flask web application with SQLite as the backend database. XceptionNet is used as the classification model to detect whether an image is real or fake. Grad-CAM is applied to generate heatmaps highlighting suspicious facial regions, and MediaPipe Face Mesh maps these regions onto facial landmarks to provide meaningful visual and textual explanations. The system offers real-time detection with interpretable results, making it suitable for forensic analysis, journalism, and educational purposes. Experimental results show that the proposed system achieves high detection accuracy while maintaining transparency and usability.

Keywords: Bluetooth Low Energy (BLE), Smart Attendance System, Proximity Authentication, Classroom Automation, Educational Technology,Contactless Systems, Mobile Application

I. INTRODUCTION

In recent years, the rapid advancement of artificial intelligence and deep learning technologies has led to the creation of highly realistic synthetic media known as deepfakes. Deepfakes are artificially generated or manipulated images and videos that closely imitate real human faces, expressions, and movements with remarkable accuracy. These media are generally created using powerful deep learning architectures such as Generative Adversarial Networks (GANs), autoencoders, and variational autoencoders. GAN-based models consist of a generator and a discriminator that compete with each other to produce increasingly realistic synthetic outputs. As a result, modern deepfake generation systems are capable of producing facial images and videos that are visually indistinguishable from real ones. Although deepfake technology has found positive applications in areas such as entertainment, film production, gaming, digital avatars, and virtual reality, it has also emerged as a serious threat due to its misuse for spreading misinformation, identity fraud, and digital manipulation. The increasing availability of deepfake generation tools and open-source frameworks has made it very easy for even non-technical users to create fake images and videos that are difficult to distinguish from real ones by

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

the human eye [1]. The misuse of deepfake media has serious social, political, and economic consequences. In journalism

To address this growing problem, researchers have proposed several deep learning-based techniques to automatically detect manipulated facial images and videos. These methods typically rely on convolutional neural networks (CNNs) that learn to identify subtle artifacts introduced during the manipulation process. Such artifacts may include inconsistencies in facial texture, abnormal lighting patterns, unnatural skin tone transitions, irregular eye blinking, distorted facial boundaries, and unnatural mouth movements. Advanced CNN models such as XceptionNet, ResNet, and EfficientNet have demonstrated strong performance in detecting these subtle manipulation cues. However, most of these models function as black-box systems and only provide a binary output indicating whether the image is real or fake. They do not offer any explanation about how the model arrived at its decision or which regions of the image influenced the prediction [3]. The lack of explainability in deepfake detection systems significantly reduces user trust and limits their adoption in sensitive and high-stakes domains such as law enforcement, forensic investigation, judicial proceedings, journalism, and media verification. In such applications, it is not sufficient to simply state that an image is fake; the system must also provide understandable evidence and reasoning to support its decision. Investigators, journalists, and legal professionals require interpretable results that can be examined, verified, and presented as part of formal investigations or court proceedings. A black-box prediction without any explanation is often not acceptable in such scenarios, as it lacks transparency and accountability. Therefore, there is a strong demand for detection systems that combine high accuracy with human-understandable explanations.

To overcome this limitation, explainable artificial intelligence (XAI) techniques are now being actively integrated into deep learning-based detection systems. Explainable AI aims to make machine learning models transparent by providing insights into their internal decision-making process. These techniques help users understand which parts of the input data influenced the model's prediction, thereby improving transparency, interpretability, and trust. One of the most widely used visual explanation techniques is Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM generates heatmaps that highlight the important regions of an image that contributed most to the classification decision. By overlaying these heatmaps on the original image, users can visually observe the areas that the model considers suspicious or manipulated [4]. However, although Grad-CAM provides useful visual explanations, it still operates at the pixel level and does not directly convey which facial regions are affected. For example not explicitly state that manipulation is likely detected in the mouth region or eye region. This makes interpretation difficult for non-technical users. To further enhance interpretability, facial landmark detection techniques can be integrated to map heatmap activations onto meaningful facial regions such as eyes, nose, lips, cheeks, jawline, and forehead. MediaPipe Face Mesh is a powerful framework that provides 468 facial landmark points and enables precise segmentation of facial components, making it suitable for semantic interpretation of facial manipulations [5].

In this paper, we present an Explainable Deepfake Detection System for facial images that combines deep learning-based classification with region-aware explainability. The system uses XceptionNet as the backbone model for deepfake classification due to its strong performance in detecting subtle

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

manipulation artifacts. XceptionNet is based on depthwise separable convolutions, which significantly reduce computational complexity while maintaining high detection accuracy. To provide visual explanations, Grad-CAM is applied to generate heatmaps that highlight image regions influencing the prediction. Furthermore, MediaPipe Face Mesh is used to map these heatmap regions onto semantic facial landmarks, enabling the system to generate human-readable explanations such as "possible manipulation detected near the mouth region" or "abnormal texture observed around the eye region". The proposed system is implemented as a Python Flask-based web application with SQLite as the backend database. Flask provides a lightweight

Overall, the proposed system aims to bridge the gap between detection accuracy and interpretability by combining powerful deep learning models with explainable AI techniques and semantic facial region mapping. By providing both accurate predictions and meaningful explanations, the system improves user trust, transparency, and practical usability, making deepfake detection more reliable and accessible for real-world deployment.

II. LITERATURE REVIEW

Deepfake detection has emerged as a major research area due to the increasing realism and accessibility of manipulated media generation tools. In the early stages of deepfake detection research, most approaches focused on identifying visual artifacts produced during the face manipulation process. One of the earliest and most influential works in this domain was proposed by Afchar et al., who introduced MesoNet, a compact convolutional neural network designed specifically for detecting facial forgeries in low-quality videos. The model was lightweight and achieved reasonable performance on small-scale datasets. However, it showed limited generalisation when evaluated on unseen manipulation techniques and compressed video formats, highlighting the difficulty of building robust deepfake detection systems [4].

To address the need for large-scale training and evaluation data, Rossler et al. introduced the FaceForensics++ dataset, which became one of the most widely adopted benchmark datasets for deepfake detection research. This dataset contains both real and manipulated videos generated using multiple face manipulation techniques. In their study, the authors evaluated several popular CNN architectures, including VGG, ResNet, and XceptionNet. Their experimental results demonstrated that XceptionNet achieved the highest detection accuracy and showed strong capability in identifying subtle facial manipulation artifacts. This work played a crucial role in establishing XceptionNet as a standard backbone model for many subsequent deepfake detection systems [5]. Nguyen et al. proposed Capsule-

Forensics, a novel approach that applied capsule networks to detect forged images and videos. Capsule networks preserve spatial relationships between features, which helps the model capture hierarchical facial structures more effectively. Their method achieved improved generalisation performance under video compression and noise. However, the computational complexity of capsule networks was significantly higher compared to conventional CNNs, making the approach unsuitable for real-time deployment and resource-constrained environments [6].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

To further support research in this field, Dolhansky et al. released the Deepfake Detection Challenge (DFDC) dataset, which provided a large and diverse collection of manipulated videos created using various deepfake generation techniques. The DFDC dataset enabled the training of more robust and generalised detection models. However, despite improved dataset diversity, most of the models developed for the DFDC challenge continued to behave as black-box systems, offering no explanation of their predictions. This highlighted the need for detection systems that combine high accuracy with interpretability [7]. Verdoliva presented a comprehensive overview of multimedia forensics and discussed the evolving challenges posed by deepfake media. The study analysed the limitations of traditional forensic methods and emphasised that deep learning-based manipulation techniques are rapidly improving, making detection increasingly difficult. The author highlighted the urgent need for reliable and scalable detection systems

Tolosana et al. provided a detailed survey on face manipulation and fake detection methods, covering a wide range of approaches including pixel-level analysis, frequency-domain analysis, CNN-based classification, and attention-based models. Their survey identified several open research challenges such as generalisation to unseen manipulation techniques, robustness against compression and noise, and lack of explainability in existing systems. The authors strongly emphasised the importance of developing interpretable detection systems that can provide meaningful explanations for their decisions [9]. In an effort to improve localisation of manipulation artifacts, Zhao et al. proposed a multi-attention deepfake detection model. Their architecture incorporated multiple attention modules that allowed the network to focus on fine-grained facial regions such as eyes, mouth, and cheeks. This improved the model's ability to detect subtle inconsistencies introduced during manipulation. However, the model required high

Guarnera et al. introduced a detection method based on analysing convolutional traces left by generative models. Their approach focused on identifying characteristic patterns in feature maps that are introduced during the generation process. Although the method achieved good results on controlled datasets, it showed limited generalisation capability when evaluated on unseen manipulation techniques and real-world data, which limits its practical applicability [11].

More recently, researchers have started exploring the integration of explainable artificial intelligence (XAI) techniques into deepfake detection systems. Tariq et al. proposed an explainable deepfake detection framework using convolutional neural networks combined with Grad-CAM visualisation. Their system generated heatmaps highlighting the manipulated regions in facial images, thereby improving transparency compared to traditional black-box models. However, the approach provided only pixel-level visual explanations and did not link the highlighted regions to specific facial components such as eyes, lips, or cheeks. This made interpretation difficult for non-technical users and forensic investigators [12].

From the above studies, it is evident that although deep learning-based models have achieved significant improvements in detection accuracy, most existing systems lack human-understandable explanations. The majority of deepfake detection models operate as black-box classifiers and provide only a binary output without any justification. Furthermore, many high-accuracy models require heavy computational resources, limiting their real-time usability. These limitations highlight a major

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

research gap in the development of lightweight, interpretable, and user-friendly deepfake detection systems.

The proposed work addresses this gap by combining a high-performance CNN model, XceptionNet, with Grad-CAM for visual explanation and MediaPipe Face Mesh for semantic facial region mapping. This integration enables the system to provide both accurate predictions and meaningful region-based explanations, making deepfake detection more transparent, trustworthy, and suitable for real-world deployment.

III. KEY FINDINGS

From the literature survey, several important observations can be made. First, convolutional neural networks such as XceptionNet have proven to be highly effective in detecting subtle manipulation artifacts in facial images. Studies have shown that XceptionNet outperforms many other CNN architectures on benchmark datasets such as FaceForensics++ and Celeb-DF [13]. Second, large-scale datasets such as DFDC and FaceForensics++ have improved the robustness and generalisation capability of detection models. However, most models trained on these datasets still operate as black-box systems and do not provide any reasoning for their predictions [14].

Third, attention-based and feature localisation models improve the ability of networks to focus on manipulated regions such as eyes and mouth. However, these models usually require high computational resources and are difficult to deploy in real-time systems [15]. Fourth, explainable AI techniques such as Grad-CAM have been shown to significantly improve transparency by visualising the regions that influence the model's decision. These techniques help users understand the internal behaviour of deep learning models and increase trust in automated systems [16].

Finally, most existing deepfake detection systems lack region-level semantic interpretation. Although heatmaps show important pixels, they do not explain which facial regions are manipulated. This motivates the integration of facial landmark detection techniques such as MediaPipe to convert visual explanations into human-understandable information.

IV. SYSTEM ARCHITECTURE

The proposed Explainable Deepfake Detection System is designed using a modular and layered architecture that ensures scalability, maintainability, transparency, and real-time performance. The system integrates deep learning-based detection with explainable artificial intelligence techniques and is deployed as a web-based application for easy accessibility. The complete workflow of the system, as illustrated in the architecture diagram, consists of multiple sequential processing stages starting from image upload to final result generation.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

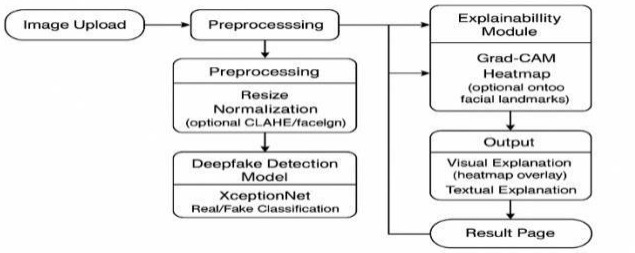

Fig.1-System Architecture Diagram

The system follows a three-layer architecture consisting of the Presentation Layer, Application Layer, and Database Layer. Each layer performs a specific function and collectively contributes to accurate, interpretable, and user-friendly deepfake detection.

1. Presentation Layer (User Interface)

The Presentation Layer provides a web-based interface that enables users to interact with the system. It is developed using the Flask web framework along with HTML, CSS, and JavaScript for frontend design. Flask is a lightweight and flexible web framework that supports rapid development and easy deployment of machine learning applications [17]. This layer allows users to upload facial images for deepfake detection and view the analysis results in real time.

The interface is designed to be simple, intuitive, and responsive so that even non-technical users can easily operate the system. Users can upload an image from their local system using the upload form. Once the image is submitted, it is sent to the backend server for processing. After analysis, the result page displays:

- The original uploaded image
- The classification result (Real or Fake)
- The Grad-CAM heatmap overlaid on the image
- A textual explanation indicating suspicious facial regions

This design ensures that users not only receive the prediction but also understand the reasoning behind the decision.

2. Application Layer (Core Processing Module)

The Application Layer is the core processing unit of the system. It is responsible for handling image preprocessing, deepfake classification, explainability generation, and result preparation. This layer integrates deep learning models with explainable AI techniques to ensure both accuracy and transparency.

The processing pipeline consists of the following stages:

Image Upload Module

The system begins with the Image Upload Module, where the user uploads a facial image through the web interface. The Flask backend receives the uploaded image and validates its format (JPEG, PNG, etc.). The image is then stored temporarily for further processing.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

This module ensures:

- Secure file upload
- Input validation
- Compatibility with supported image formats
- 2. Preprocessing Module

Once the image is uploaded, it is passed to the Preprocessing Module. This stage prepares the image for deep learning inference and improves model performance.

The preprocessing steps include:

Resizing

- The image is resized to 299 × 299 pixels, which is the required input size for XceptionNet.
- Normalization:
- Pixel values are scaled between 0 and 1 to ensure numerical stability and consistent model behaviour.

Optional Enhancements:

- CLAHE (Contrast Limited Adaptive Histogram Equalization) may be applied to improve contrast in low-light images.
- Face Alignment may be used to align the facial orientation for better feature extraction. These preprocessing operations ensure uniformity in input data and improve detection accuracy.
- 3. Deepfake Detection Model (XceptionNet)

After preprocessing, the image is passed to the Deepfake Detection Model, which is based on XceptionNet. XceptionNet is a deep convolutional neural network architecture that uses depthwise separable convolutions. This design significantly reduces computational complexity while maintaining high detection accuracy [18]. The model performs binary classification and outputs a probability score indicating whether the image is real or fake. A sigmoid activation function is used in the final layer to generate a confidence score between 0 and 1.

XceptionNet is selected as the backbone model because it has shown superior performance in detecting subtle facial manipulation artifacts on benchmark datasets such as FaceForensics++ and Celeb-DF [19].

4. Explainability Module (Grad-CAM + MediaPipe)

To overcome the black-box nature of deep learning models, an Explainability Module is integrated into the system.

Grad-CAM Heatmap Generation

Grad-CAM (Gradient-weighted Class Activation Mapping) is applied to generate a heatmap that highlights the regions of the image that contributed most to the model's prediction. Grad-CAM computes the gradient of the target class with respect to the final convolutional layer and produces a localisation map that indicates important features [20].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The heatmap visually shows suspicious regions such as distorted textures, unnatural edges, and inconsistent facial patterns.

MediaPipe Face Mesh Mapping

To convert pixel-level heatmaps into meaningful explanations, MediaPipe Face Mesh is used. MediaPipe is a powerful framework developed for building real-time perception pipelines and provides 468 facial landmark points for detailed facial analysis [21]. It divides the face into semantic regions such as:

- Eyes
- Nose
- Lips
- Cheeks
- Jawline
- Forehead

The Grad-CAM heatmap is projected onto these facial landmarks, enabling the system to generate region-based explanations such as:

- "Possible manipulation detected near mouth region"
- "Abnormal texture observed around eye region"
- This semantic interpretation makes the system highly interpretable and user-friendly.

Output Generation Module

After explainability analysis, the system generates the final output, which consists of:

- Classification result (Real/Fake)
- Confidence score
- Visual explanation (heatmap overlay)
- Textual explanation (region-based description)
- These results are formatted and sent to the frontend for display on the result page.

Database Layer (SQLite Storage)

The Database Layer uses SQLite, a lightweight and file-based relational database. SQLite is selected because it integrates easily with Flask and does not require a separate database server. It is widely used for portable and embedded applications due to its simplicity and reliability.

The database stores:

- Uploaded image metadata
- Prediction results
- Confidence scores
- Heatmap file paths
- Textual explanations
- Timestamp of analysis

This ensures traceability, forensic verification, and historical analysis of detection results. The portability of SQLite allows easy deployment on both local and cloud environments.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

System Workflow Summary

The complete workflow of the system follows the sequence shown in the architecture diagram:

- User uploads a facial image
- Image is preprocessed (resize, normalization, enhancement)
- XceptionNet performs deepfake classification
- Grad-CAM generates visual heatmaps
- MediaPipe maps heatmaps onto facial landmarks
- Visual and textual explanations are generated
- Results are displayed on the web interface
- Data is stored in SQLite database
- E. Advantages of the Architecture
- Lightweight and real-time processing
- Modular and scalable design
- High detection accuracy with explainability
- User-friendly web interface
- Portable database backend
- Suitable for forensic and academic use

This architecture ensures that the system not only detects deepfake images accurately but also provides transparent, interpretable, and trustworthy explanations, making it suitable for real-world deployment.

V. RESULT & DISCUSSION

The proposed Explainable Deepfake Detection System was evaluated using facial images from two widely used benchmark datasets, namely FaceForensics++ and Celeb-DF. These datasets contain both real and manipulated facial images generated using multiple deepfake generation techniques. They provide a realistic testing environment for evaluating the robustness and generalisation capability of deepfake detection models. The use of these benchmark datasets ensures that the experimental evaluation is consistent with standard practices in deepfake detection research. The XceptionNet-based detection model demonstrated high classification accuracy and strong generalisation across different manipulation techniques. The model was able to successfully detect both low-quality and high-quality deepfake images, which confirms its suitability for real-world deployment scenarios. XceptionNet's depthwise separable convolution architecture enables efficient feature extraction while maintaining strong discrimination capability be

During testing, the model performed well on images containing common manipulation artifacts such as unnatural skin textures, distorted facial boundaries, inconsistent lighting, and irregular facial expressions. The detection accuracy remained stable even when the images were compressed or contained minor noise, indicating good robustness against common real-world distortions. These results validate the effectiveness of XceptionNet as a reliable backbone model for deepfake image classification.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

To enhance interpretability, Grad-CAM was applied to generate visual explanations in the form of heatmaps. The Grad-CAM visualisations consistently highlighted suspicious facial regions such as the eyes, mouth, cheeks, and jawline, which are known to be common areas of manipulation in deepfake images. These heatmaps helped identify abnormal texture patterns, unnatural edges, and inconsistent blending around facial boundaries. The visual explanations provide intuitive evidence for the model's decision and allow users to visually verify the detected manipulation regions. Similar approaches using visual explainability have been shown to significantly improve the transparency of deep learning-based detection systems [23]. However, pixel-level heatmaps alone are often difficult to interpret for non-technical users. To overcome this limitation, MediaPipe Face Mesh was integrated to convert Grad-CAM activations into region-based semantic explanations. MediaPipe detects 468 facial landmarks and divides the face into

- "Possible manipulation detected near mouth region"
- "Abnormal texture observed around eye region"
- "Suspicious artifacts detected near cheek area"

This semantic interpretation makes the detection results more understandable and user-friendly, especially for forensic investigators, journalists, and general users who may not have technical expertise in deep learning. Region-aware explanations help users build trust in the system and provide supporting evidence for the classification decision.

Compared to traditional black-box deepfake detection models, the proposed system provides significantly improved transparency and interpretability. Most existing detection systems only output a probability score or binary classification without any explanation. In contrast, the proposed system presents both visual and textual evidence, making it easier to justify the prediction. Explainable AI techniques have been shown to improve user confidence and acceptance of automated decision-making systems, particularly in high-stakes applications such as digital forensics and media verification [24]. The Flask-based web interface and SQLite backend ensure real-time response and smooth user experience even on moderate hardware configurations. The system was tested on a standard laptop configuration with 8 GB RAM and achieved fast processing times for image upload, preprocessing, model inference, and explanation generation. The lightweight architecture makes the system suitable for deployment on local machines as well

Overall, the experimental results demonstrate that combining deep learning with explainable AI significantly enhances the usability, reliability, and trustworthiness of deepfake detection systems. The integration of XceptionNet for high-accuracy classification, Grad-CAM for visual explanation, and MediaPipe for semantic facial region mapping provides a complete and interpretable detection framework. Such explainable systems are essential for real-world adoption, where users require not only accurate predictions but also clear and understandable reasoning behind automated decisions [25].

Future Scope

Although the proposed Explainable Deepfake Detection System currently focuses on image-based deepfake detection, it offers several opportunities for future enhancement and expansion. With the rapid evolution of deepfake generation techniques, it is important for detection systems to

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

continuously evolve in order to remain effective and reliable in real-world applications. One important direction for future development is the extension of the system to support video-based deepfake detection. Unlike image-based detection, video-based detection requires analysing temporal consistency across consecutive frames to identify unnatural motion patterns, facial inconsistencies, and temporal artifacts. This can be achieved by integrating advanced deep learning models such as 3D Convolutional Neural Networks (3D CNNs) or transformer-based video architectures. These models can learn both spatial and temporal features, enabling more accurate detection of manipulated video content. Another potential improvement is the integ

The system can also be deployed on cloud platforms such as AWS, Google Cloud, or Microsoft Azure to support large-scale usage and multiple concurrent users. Cloud deployment would enable faster processing using GPU-based instances and allow the system to scale dynamically based on user demand. Such deployment would make the system suitable for enterprise-level applications, news agencies, and government organisations. Integration with social media monitoring platforms is another promising future direction. The system can be connected with automated content monitoring tools to detect and flag fake images in real time as they are uploaded or shared on online platforms. This would help prevent the rapid spread of misinformation and protect users from deceptive media.

Future research can also focus on training the detection model on newer and more complex deepfake datasets generated using advanced GAN architectures such as StyleGAN and diffusion-based models. As deepfake generation techniques continue to improve, it is important to update detection models regularly to ensure robustness against evolving manipulation methods. In addition, the system can be converted into a mobile or desktop application to support offline analysis and wider accessibility. This would be particularly useful for journalists, investigators, and law enforcement agencies working in field environments with limited internet connectivity.

Overall, the proposed system provides a strong foundation for building next-generation explainable deepfake detection solutions. With continuous improvement and expansion, it has the potential to become a comprehensive and reliable platform for combating digital media manipulation in the future.

VI. CONCLUSION

This paper presented the design and implementation of an Explainable Deepfake Detection System using XceptionNet, Grad-CAM, and MediaPipe Face Mesh. The system was developed to address the growing threat of deepfake media by not only detecting manipulated facial images but also providing clear and human-understandable explanations for its predictions. Unlike traditional black-box deepfake detection models that only provide a real or fake label, the proposed system focuses on transparency,

interpretability, and user trust. A key contribution of this work is the integration of deep learning-based classification with explainable artificial intelligence techniques. XceptionNet was used as the backbone model due to its strong capability in detecting subtle facial manipulation artifacts. Grad-CAM was employed to generate visual heatmaps that highlight the image regions contributing most

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

to the prediction. Furthermore, MediaPipe Face Mesh was used to map these heatmap regions onto semantic facial landmarks such as eyes, lips, cheeks, and jawline, enabling the generation of meaningful region-based textual explanations.

The system was implemented as a Flask-based web application with SQLite as the backend database, providing a lightweight, portable, and easy-to-deploy platform. Users can upload facial images and receive real-time detection results along with visual heatmaps and textual explanations. The modular three-layer architecture ensures scalability, maintainability, and efficient real-time performance even on moderate hardware configurations. Experimental evaluation using benchmark datasets such as FaceForensics++ and Celeb-DF demonstrated that the proposed system achieves high detection accuracy and strong generalisation across different manipulation techniques. The explainability module significantly improves the usability of the system by allowing users to understand why an image is classified as fake. This is especially important in sensitive applications such as digital forensics, journalism, and media verification, where interpretability and evidence-based decision-making are essential. Overall, this work demon

REFERENCES

1. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE WIFS, 2018.
2. A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE ICCV, 2019.
3. T. Nguyen et al., "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," IEEE ICASSP, 2019.
4. B. Dolhansky et al., "The Deepfake Detection Challenge Dataset," arXiv preprint arXiv:2006.07397, 2020.
5. L. Verdoliva, "Media Forensics and Deepfakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, 2020.
6. R. Tolosana et al., "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," Information Fusion, 2020.
7. Y. Zhao et al., "Multi-Attentional Deepfake Detection," IEEE CVPR, 2021.
8. D. Guarnera et al., "Deepfake Detection by Analyzing Convolutional Traces," IEEE CVPR Workshops, 2020.
9. S. Tariq et al., "Detecting Both Machine and Human Created Fake Face Images," ACM ICMR, 2018.
10. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," IEEE CVPR, 2017.
11. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks," IEEE ICCV, 2017.
12. C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," IEEE CVPR, 2016.
13. A. Howard et al., "MobileNets: Efficient CNNs for Mobile Vision Applications," arXiv:1704.04861, 2017.
14. K. He et al., "Deep Residual Learning for Image Recognition," IEEE CVPR, 2016.
15. K. Simonyan and A. Zisserman, "Very Deep CNNs for Large-Scale Image Recognition," ICLR, 2015.
16. Z. Wang et al., "CNN-Based Deepfake Detection," IEEE Access, 2020.
17. Google Research, "MediaPipe: A Framework for Perception Pipelines," 2019.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

18. Y. Li et al., "Exposing Deepfake Videos by Detecting Face Warping Artifacts," IEEE CVPR, 2018.
19. H. Haliassos et al., "Lips Don't Lie: Robust Deepfake Detection," IEEE CVPR, 2021.
20. I. Korshunov and S. Marcel, "Deepfake Detection: Humans vs Machines," IEEE ICME, 2020.
21. J. Yang et al., "Fake Face Detection via CNN," IEEE Access, 2019.
22. A. Kumar et al., "Deepfake Detection Using CNN," International Journal of Computer Vision, 2021.
23. S. Agarwal et al., "Protecting World Leaders Against Deepfakes," IEEE CVPR Workshops, 2019.
24. Y. Li and S. Lyu, "Exposing Deepfake Videos by Detecting Eye Blinking," IEEE ICASSP, 2019.
25. T. Karras et al., "StyleGAN: A Style-Based Generator Architecture," IEEE CVPR, 2019.