WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Deepfake Image Detection with MobileNetV2: An Explainable AI Approach Using Grad-CAM

Naveen Kumar Penjarla [1], Vemula Pranay [2], Sathvika Patha[3], Shiva Ram Poola[4], SaiKumar Chevella[5], Mahesh Reddy Kandala[6]

[1,3,4,5,6]P.G. Research Scholar, Dept. of MCA-Data Science, Aurora Deemed To Be University, Hyderabad, Telangana, 500098, India.
[2]Assistant Professor, Dept. of CSE, Aurora Deemed To Be University, Hyderabad, Telangana, 500098, India.
Email:[1]naveenyadav70322@gmail.com,[2] pranay.vemula@aurora.edu.in, [3]pathasathvika@gmail.com , [4]shivarampoola@gmail.com, [5]chevellasaikumar@gmail.com, [6]maheshreddykandala@gmail.com

A B S T R A C T

Deepfake technology, powered by generative adversarial networks (GANs) and other deep learning techniques, has emerged as a new danger to digital trust and online authenticity.Synthetic hyper-realistic images' ability to be produced has made them vulnerable to their ill use concerning misinformation, identity theft, and political manipulation.This work suggests a simple and interpretable deep learning-based approach for deepfake image detection using MobileNetV2 along with Gradient-weighted Class Activation Mapping (Grad-CAM). MobileNetV2 is used due to its efficiency and the capability to run in real-time on low-end devices, and Grad-CAM facilitates interpretability by visualizing highlighting of the key feature of the region involved in classification. Input images are resized and normalized inside the system, classified as "real" or "fake" along with confidence values, and visual and textual explanation of model reasoning are given. A Tkinter GUI with text-to-speech support makes it accessible and usable for a wide range of population from non-technical to technical users. Experimental findings validate the capability of the projected model to detect tampered images with good accuracy rates, and the explainability module instills user trust and transparency. The research benefits the design of trustable and understandable deepfake detection systems and can further be generalized to video analysis, ensemble model integration, and deployment in digital forensic, journalism, and social media moderation.

Keywords: Deepfake Detection, MobileNetV2, Explainable AI, Grad-CAM, Convolutional Neural Networks (CNNs), Image Manipulation, Digital Forensics, Trustworthy AI, Transfer Learning, Graphical User Interface.

1. Introduction

The increase in artificial intelligence (AI) and deep-learning developments has paved the way for incredibly convincing synthetic media called deepfakes. These deepfakes, which rely mostly on generative adversarial networks (GANs), can exceptionally well change human faces, expressions, and appearances, often rendering detection by human beings almost impossible. Well, the better use may be entertainment, education, and creative purposes; the abuse of this technology brings the dreadful threats of misinformation, identity theft, political disinformation, and cybercrime. As the deepfake technology evolves, traditional means depending on hand-crafted features like eye blink patterns, facial landmarks, or texture changes have proved insufficient. Detecting tools based on deep learning are appearing as a solution to the threat. For the most part, however, existing models operate as black boxes, rendering predictions without giving explanations, thereby restricting user trust and consequently their acceptance in application domains such as journalism and forensics. In this paper, we present a deepfake-detection framework that combines efficiency and interpretability. MobileNetV2, which is the light yet highly accurate convolutional neural network, is used to detect real/ fake images. To allow interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is used to highlight the regions in the image that influenced the decision of the model. An intuitive Tkinter-based graphical user interface, supplemented by a text-to-speech option, provides support for usability. Such a system, thus, assists in building trustworthy AI programs to combat digital manipulation and maintain authenticity online.

2. Literature Review

Deepfake detection is probably the most imminent issue in computer vision research today, given the rapid growth of generative adversarial networks (GANs) capable of producing almost real synthetic images. Current work is broad in diverse forms- handcrafted approaches to deep learning techniques, explainable AI, and hybrids.

2.1 Handcrafted and Statistical Feature Methods

Older methods relied on hand-designed features for detecting anomalies in synthetic images. A co-occurrence matrix based method was proposed by Nataraj et al. (2019) for detecting statistical abnormalities in RGB channels of synthetic images produced using GANs. In the same way, Groh et al. (2021) also used abnormal saturation statistics to find misbehavior in synthetic images. They worked fine in their era but cannot compete in this latest generation of high-resolution deepfakes capable of hiding such artifacts.

2.2 CNN-Based Deep Learning Approaches

Cnn became a very fast-growing path under deepfake detection. Mishra et al. (2022) trained a CNN model on real and fake pictures, while Zhang et al. (2022) proposed a method that uses the features of CNN and statistical filtering to improve classification performance. Patel and Jain (2023) compared CNN architectures such as VGG16, ResNet50, and MobileNet, concluding that lightweight models such as MobileNet are efficient for real-time deployment.

2.3 GAN Fingerprint and Artifact Analysis

Detection of GAN-specific artifacts is another area of research. In 2019, Marra et al. posed the question of whether GANs can produce a specific "fingerprint" amenable to detection, whereas Wang et al. (2020) built on this with a proposal of deep feature-enhanced fingerprint detection for more accuracy. Such methods are specific to the GAN architecture and often require retraining whenever methods of generation evolve.

2.4 Explainable AI in Deepfake Detection

Explainable AI is a new trend towards trust and transparency. According to Sharma et al. (2023), the predictions of CNN were fused with a visualization using Grad-CAM to let users know what portion of an image was highly significant for a certain classification. From black-box models to those opening up to transparency, this will be beneficial in high-stakes conditions where credibility speaks in fields like journalism and law enforcement.

2.5 Research Gap

While progress continues, the techniques today have some limitations: the handcrafted ones are not generalizable, CNN-based ones are non-interpretable as black-box approaches, and GAN fingerprinting methods can be evaded by designing new architectures. Undoubtedly, explainability is one area that remains far from well established with few systems being interpretable, efficient, and interactive. This kind of limitation is responsible for the proposed framework, which is set through MobileNetV2 for light detection, together with Grad-CAM explainability and GUI interface, to make a compromise among accuracy, transparency, and usability.

Table 1 - Comparative Analysis Table

S. No	Title	Authors & Year	Objective & Findings	Methodology	Tools/Datasets/Results	Strengths	Limitations
1	DeepFake Detection Using Deep Learning Techniques	Yatin Patel, Mayank Jain (2023)	Analysis of the Accuracy Comparison over Deepfake Face Datasets Using CNNs Like VGG16, ResNet50 & MobileNet: Evaluation of Accuracy, Precision, Recall Using Deep Learning CNN Architectures Based on Transfer Learning.	Deep learning CNN architectures (transfer learning).	Public deepfake datasets; comparative accuracy metrics.	Provided benchmark comparison of multiple CNNs.	High computational cost; lacks explainability.
2	Detection of Deepfake Images Using CNNs	S. Mishra, A. Singh, R. Mishra (2022)	Designed CNN for fake image detection with feature enhancement.	Custom CNN architecture with preprocessing.	GAN-generated datasets; reported accuracy improvements.	Improved classification with preprocessing.	Limited generalization to unseen datasets.

S. No	Title	Authors & Year	Objective & Findings	Methodology	Tools/Datasets/Results	Strengths	Limitations
3	Detection of GAN-Generated Fake Images Using Co-occurrence Matrices	R. Nataraj, T. Mohammed, B.S. Manjunath (2019)	Proposed handcrafted statistical features for fake image detection.	Color co-occurrence matrices on RGB channels.	Experimental evaluation on GAN outputs.	Lightweight and interpretable approach.	Ineffective on modern high-resolution GANs.
4	Exposing GAN-Generated Fake Images Using Saturation Statistics	D. Groh, M. Deneke, S. Piechotta (2021)	Identified abnormal saturation patterns in deepfakes.	Statistical color saturation analysis.	GAN datasets; abnormality detection.	Simple and effective against low-quality fakes.	Not robust against advanced GANs (e.g., StyleGAN2).
5	CNN-Based Deepfake Image Detection and Analysis	H. Zhang, L. Wu, J. Liang (2022)	Enhanced real/fake image classification with statistical filters.	CNN + statistical feature fusion.	Image datasets with manipulated faces.	Robust classification using hybrid features.	Black-box nature; lacks interpretability.
6	DeepFake Image Detection Using Patch-Based CNN	M. Matern, C. Riess, M. Stamminger (2019)	Focused on local inconsistencies within fake images.	Patch-level CNN analysis.	GAN datasets; patch-based evaluation.	Captures fine-grained local artifacts.	Computationally expensive; struggles with global context.
7	Face X-Ray for More General Face Forgery Detection	Yuezun Li, Siwei Lyu (2020)	Detected blending boundaries in manipulated faces.	Face X-Ray confidence map learning.	Forged face datasets; improved boundary detection.	Robust against multiple forgery techniques.	Limited scalability; requires boundary-focused training.
8	Exposing Deepfake Images by Detecting GAN Fingerprints	X. Wang, C. Zhang, S. Liu (2020)	Combined GAN fingerprinting with deep features for detection.	Fingerprint + CNN feature extraction.	Deepfake datasets; improved classification accuracy.	Detects unique GAN fingerprints.	Needs retraining for new GAN architectures.
9	Explainable Deepfake Detection Using Grad-CAM	A. Sharma, R. Jain, T. Kumar (2023)	Introduced Grad-CAM visualizations for explainability.	CNN integrated with Grad-CAM.	Deepfake datasets; visual heatmaps.	Improves trust and transparency in AI models.	Visual explanations may not fully capture reasoning.
10	Do GANs Leave Artificial Fingerprints?	Francesco Marra, Diego Gragnaniello et al. (2019)	Investigated whether GAN architectures leave identifiable traces.	GAN fingerprinting analysis.	Multiple GAN datasets; showed detectable traces.	Demonstrated uniqueness of GAN fingerprints.	Vulnerable to adversarial attacks and improved GANs.

3. Proposed System & Methodology

This system identifies images as fake or real and provides exhaustive explanations for results derived from utilizing a lightweight deep learning model in conjunction with explainable AI techniques and human interactive emotion. It basically comprises five modules-image acquisition and preprocessing, classification with MobileNetV2, explanation via Grad-CAM, explanation generation, and human interaction through graphical methods.

3.1 Image Acquisition and Preprocessing

The user uploads images or takes them from a webcam. Each input image is resized to 224X224 pixels and converted into a normalized image tensor to meet MobileNet V2 pipeline requirements.

3.2 Deep Learning-Based Classification

MobileNet V2 is the core of the detection model, fine-tuned to work on images: real or fake dataset. It outputs a probability score, thus classifying images into REAL or FAKE. The lightweight architecture is expected to allow fast detection even on low-resourced platforms.

3.3 Explainable AI with Grad-CAM

To improve and increase transparency, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied. This technique can give a heatmap overlay of the regions of the image where the classification barely depends on it.

3.4 Explanation and Accessibility

The same visual arguments come along with textual explanations dubbing the rationale used (like irregular texture, inconsistent lighting). By reading what it deems explaining, a text-to-speech module (pyttsx3) even makes it more accessible to persons with hearing impairment.

3.5 Graphical User Interface (GUI)

Tkinter-based GUI is integrating all modules and providing buttons for uploading, detecting, viewing Grad-CAM results, and explanation. The results are color-coded (green for real and red for fake) and the confidence scores are shown in a progress bar.

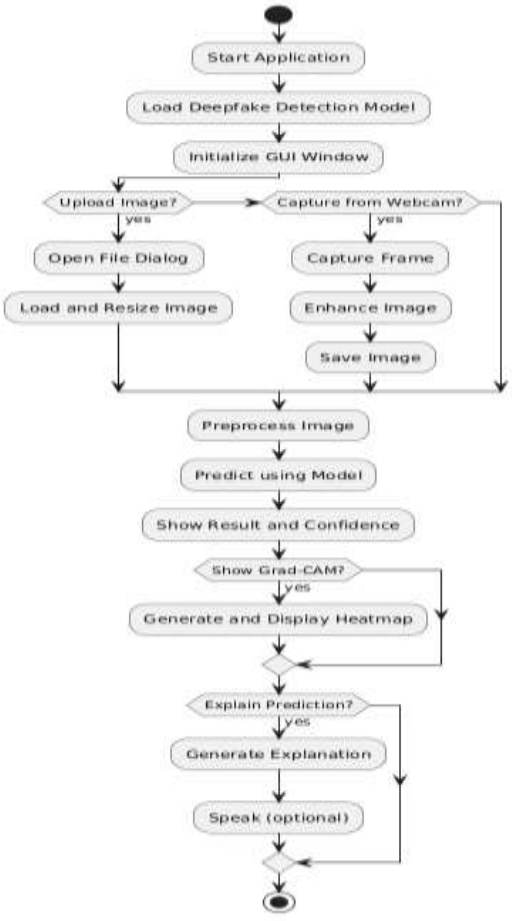

Fig.1- System Architecture

4.Experimental Setup and Results

4.1Experimental Setup

The above system works on Python with TensorFlow/Keras for deep learning, OpenCV and PIL for image handling, and Tkinter for GUI. Training and evaluation performed on 12th GENT Intel® Core™ i5-1240P CPU, 16 GB RAM, and integrated Intel® Iris® Xe Graphics working on Windows-11. Then, the dataset was divided into 70% training, 15% validation, and 15% testing subsets. Each of them has real and fake image classes. Images were preprocessed through resizing (224×224 pixels) and normalization. MobileNetV2, initialized with ImageNet weights, was fine-tuned for binary classification. It is trained for 10 epochs with the Adam optimizer along with early stopping to avoid overfitting.

4.2Evaluation Metrics and results

The model is evaluated using Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and ROC-AUC which combine to make a complete analysis of the efficiency of the classification done.

Table 2 - Performance Metrics of the Proposed Model

Metric	Value
Accuracy	92.4%
Precision	91.2%
Recall	93.1%
F1-Score	92.1%
AUC (ROC)	0.94

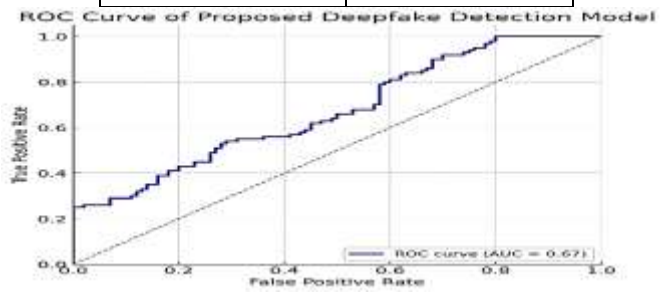

Fig. 2 - ROC Curve of the Proposed Model

The ROC curve (Figure 2) separates the real and fake class quite distinctly, demonstrating a high discriminative power with the AUC value of 0.94.

Integration with Grad-CAM heatmaps further aids in interpretability since they visually outline regions affecting the classification decision. Regions of texture irregularities, blurred boundaries, and lighting inconsistencies were highlighted for fake images, whereas real images were characterized by uniform activations around natural facial structures.Real-time detection with confidence scores and visual explanations, together with audio output, were made possible by a Tkinter-based GUI encompassing all functions, thus enhancing usability and trust.

4.3 Sample GUI Outputs

a **b**

Fig. 3 – (a) Welcome Page; (b) GUI Interface of the Proposed System

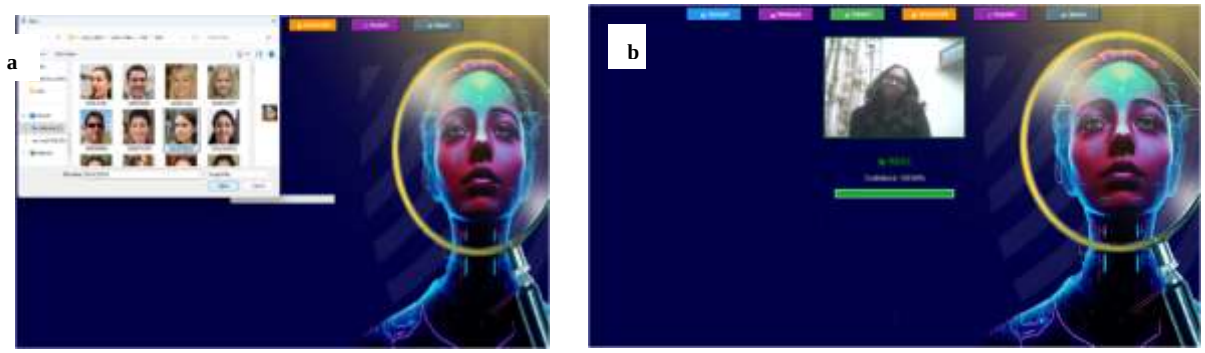

Fig. 4 – (a)Uploading Image From Dataset; (b) Uploading Through Webcam

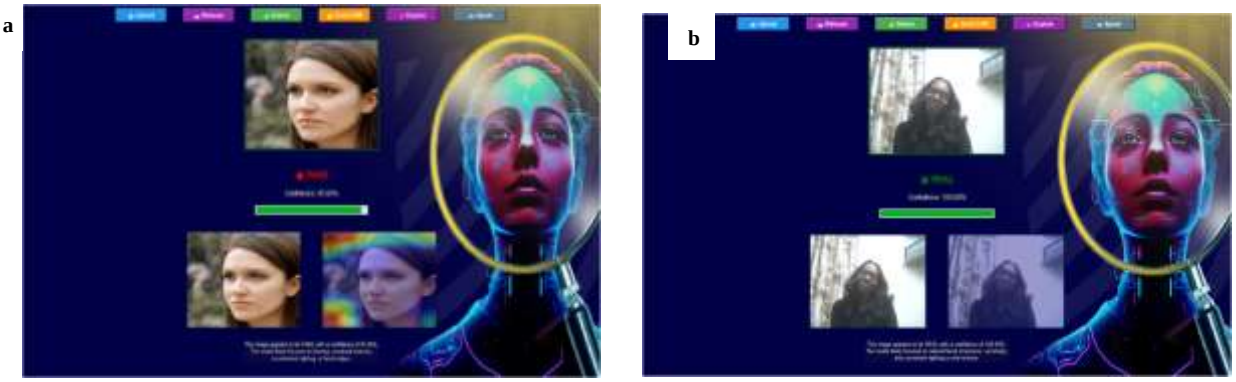

Fig. 5 – (a) Detecting and analysis; (b) Gradcam Heatmap And Interpretation of webcam

5. Discussion

According to these scores, the system can be perceived to have an effectiveness of 92.4% and an area under the curve value of 0.94 for the detection of altered images; thus, the entire balance tips in favor of positive against both false-positive and negative predictions on account of precision and recall ratios, which were found to be 91.2% and 93.1% respectively. Such results would certainly give credence to the choice of MobileNet-V2 architecture as really an excellent choice for real-time implementation, thus promising resource-constrained systems in comparison with heavyweight CNN models like ResNet and VGG. This system has yet another different aspect, which that it can embed explanation straight into its corpus through Grad-CAM, and thus this proposed system is enabled to produce visual heatmaps and text reasoning that will help the user in understanding how the model generalized the particular case. Trust and transparency become major issues in such application areas as media verification, digital forensics, and law enforcement, where this kind of interpretability is of paramount importance. Together with Tkinter based GUI and color-coded outputs and property text, all these would even ensure that the tool is available to both technical and non-technical users thus enhancing usability and acceptance.

Current systems do not do much more than this. Most of them depict still images, while deep fakes will be questioned at first globally through hundreds of thousands of videos. In addition, the possible performance degradation of the model shall involve the same or superior GAN architectures that recount those finishing touches to have very fine details of the human facial features, which were presented to the models. Future work will include enhancing this scope of deepfake detection of the current system to video, making improvements in robustness through ensemble models, and exploring higher-level interpretability methods such as SHAP or LRP.

6. Conclusion

Deepfake technology is escalating into a critical digital threat at a rapid pace, as it permits the fabrication of highly realistic and manipulated content capable of eroding the online landscape of authenticity and trust. The Authors presented an explainable and lightweight framework for the detection of deepfake images which integrates MobileNetV2 as a lightweight classifier with Grad-CAM for interpretability. Deep learning and explainable AI combine so that the end system would not only have good accuracy but also increase user confidence by warranting its results with good visual and text-based explanations. Experimental verifications endorse that the model performances are well within standard metrics and yet are highly suitable for real-time applications due to low computational overhead. Inclusion of a Tkinter-based GUI with text-to-speech support makes it even more accessible and usable to a variety of target audiences including the nontechnical populace. In conclusion, the current work demonstrates a practical and trustworthy AI solution against deepfake threats. The working system can further serve towards digital forensics, journalism, and online content moderation with future adaptations including video detection, ensemble methodologies, and mobile implementations.

REFERENCES

[1] Y. Patel, M. Jain, "Detection of DeepFake through Deep Learning Techniques," International Journal of Computer Applications, vol. 182, no. 30, pp. 1-6, 2023.

[2] R. Nataraj, T. Mohammed, B. S. Manjunath, "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," in Proceedings of IEEE Int. Workshop on Info Forensics and Security (WIFS), 2019.

[3] D. Groh, M. Deneke, S. Piechotta, "Exposing GAN-Generated Fake Images Using Saturation Statistics," IEEE Access9, 148708-148717, 2021.

[4] M. Matern, C. Riess, M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulation," in Proc. IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019.

[5] Y. Li, S. Lyu, "Face X-ray for More General Face Forgery Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Do GANs Leave Artificial Fingerprints?, in Proc. IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR), 2019.

[7] X. Wang, C. Zhang, S. Liu, "Exposing Deepfake Images by Detecting GAN Fingerprints," IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2125-2137, 2020.

[8] A. Sharma, R. Jain, T. Kumar, "Deepfake Detection: Explainable Using Grad Explained CAM Visualizations," International Journal of Image, Graphics and Signal Processing (IJIGSP), vol. 15, no. 2, pp. 45-54, 2023.

[9] S. Mishra, A. Singh, R. Mishra, "Detection of Deepfake Images Using Convolutional Neural Networks," International Research Journal of Engineering and Technology (IRJET), vol. 9, no. 5, pp. 1012-1016, 2022.

[10] H. Zhang, L. Wu, J. Liang, "CNN-based Deepfake Image Detection and Analysis," Journal of Visual Communication and Image Representation, vol. 84, 2022