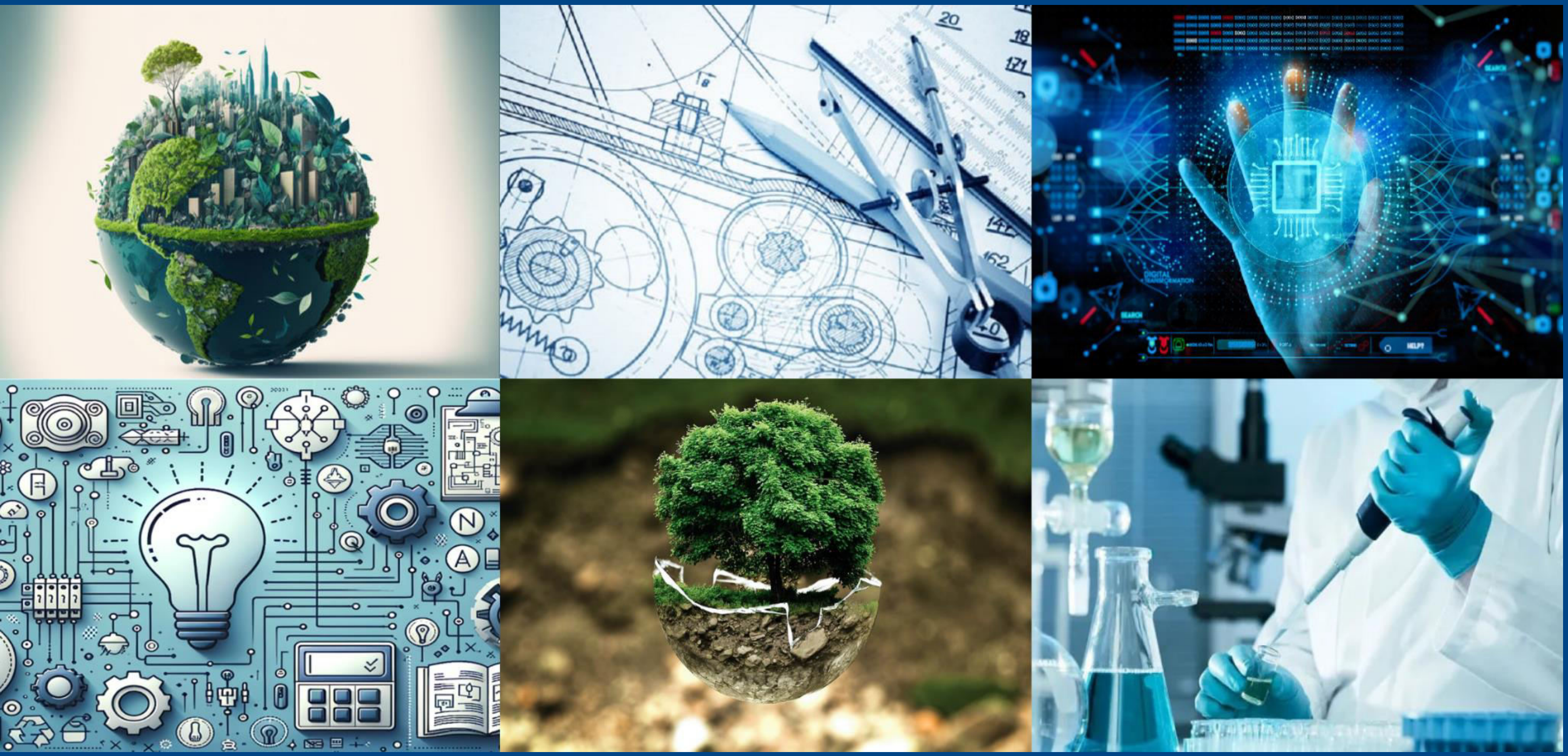
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Issue 6, June 2025

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806197

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

DeepGuard AI: Real-Time DeepFake Detection

Swathi Gowroju[1], Kaushik[2], Harshith Reddy[3], Vishnupriya [4], Krishna Teja[5]

Associate Professor, Dept. of CSE (AI&ML), Sreyas Institute of Engineering and Technology, Telangana, India[1]

Dept. of CSE (AI&ML), Sreyas Institute of Engineering and Technology, Telangana, India[2]

Dept. of CSE (AI&ML), Sreyas Institute of Engineering and Technology, Telangana, India[3]

Dept. of CSE (AI&ML), Sreyas Institute of Engineering and Technology, Telangana, India[4]

Dept. of CSE (AI&ML), Sreyas Institute of Engineering and Technology, Telangana, India[5]

ABSTRACT: Given the rise of artificial intelligence and generative media technologies, deepfake technology threatens our digital trust, public safety, and online security. AI-enabled false photos and videos can prove hard to forensically identify, which can lead to misinformation, identity theft, and reputational harm spreading rapidly through our communities. We propose one solution, DeepGuard AI, which will provide a real-time system to detect deepfake photo material and clearly label it so it cannot be exploited.

DeepGuard AI provides an ensemble learning model consisting of three different deep learning models, EfficientNetB0, Xception, and ResNet50, to evaluate facial characteristics and classify all input images as either real or fake with accurate predictive confidence. When deepfake material is detected by the system, the original image can receive a watermark that states "DEEPFAKE" deep into the material, visually depicting the tampering. This approach provides a two-gate method that both detects and prevents any future misuse and distribution of the original image. This system is built in a Flask-based API that accepts two forms of input: image URLs or base64-encoded image data. The system returns "live" predictions that include the expected label (real or fake), the confidence score, and the watermarked image (if it was fake). These features allow this designed system to provide easily accessible and useful functionality.

KEYWORDS: Deepfake Detection, Artificial Intelligence, Ensemble Learning, EfficientNetB0, Xception, ResNet50, Tamper-Proof Watermarking

I. INTRODUCTION

The advancement of artificial intelligence and deep learning has activated the creation of exciting forms of synthetic media called deepfakes. Deepfakes are AI-generated images and videos that can mimic real people (and other entities) with impressive resemblance and fidelity. The rise of deepfakes has continued to set in motion related concerns about misinformation, privacy violations, identity theft, and security risks. As a result, deepfakes have become the state of the art, and many are increasingly being used with discretion. The market for clever tools to help determine real from fake media is on the increase. This project represents a system to address these demands with an immediate and pragmatic solution—enabling real-time deepfake media detection and additionally watermarking tamper-proof "DEEPFAKE" labels on images of subjects - effectively stopping any pathways to spreading or using incorrect or pernicious deepfakes otherwise. For high accuracy, The proposed system is an ensemble deep learning architecture comprised of three strong CNN-based architectures: EfficientNetB0, Xception, and ResNet50 - all of someone else's work fine-tuned through a deepfake dataset we compressed on Kaggle.com. Media input can take place through image URL links or base64 data to help with portability, usable in web apps for forensic tools or social media.

II. LITERATURE SURVEY

While deepfakes are on the rise, they have also led to a wave of research on media detection. At an early time (2019), Korshunov and Marcel [1] proposed that deepfakes are a threat to face recognition systems and that we must develop detection methods urgently. Later, Li and Lyu [2] pointed out that detecting the arc of face-warping artifacts is an early and effective lead in deepfake videos. Afchar et al. [3] presented MesoNet, a new lightweight CNN architecture that was designed specifically for deepfake detection. MesoNet used mesoscopic features while outperforming conventional forensic methods. Rössler et al. [4] proposed the FaceForensics++ Benchmark dataset, a widely used dataset for deepfake detection researchers to further develop models. Mittal et al. [5] added some appearance features and subtle features like eye blinking and head tilting to assist in video-based detection. At the model architectural level, Xception [6], EfficientNet [7], and ResNet [8] have proven useful for image classification tasks for deepfake detection. The deepfake networks above form the basis of DGA, along with additional deepfake detection systems. There has also been research that evaluates wider and more general anti-spoofing methods. Wang et al. [9] shared a position paper that provided robust insights into the model.

Table I: Discussion about Various Related Projects

Reference	**Author**	**Key Focus**	**Methodology**	**Results**
[1]	Korshunov & Marcel (2018)	Deepfake risks to face recognition systems	Experimental study with biometric systems	Identified major security risks and detection gaps
[2]	Li & Lyu (2019)	Detecting face warping artifacts	Pixel-level artifact analysis	Effective in early deepfake detection
[3]	Afchar et al. (2018)	Compact CNN for forgery detection	Proposed MesoNet CNN	Achieved ~82% accuracy; lightweight model
[4]	Rössler et al.(2019)	Dataset for training/testing deepfake models	Created FaceForensics++ dataset	Became a benchmark in deepfake research
[5]	Mittal et al. (2020)	Detecting deepfakes via appearance + behavior	Behavioral + visual analysis	Improved accuracy in video-based fakes
[6]	Chollet (2017)	Depthwise separable CNN for better learning	Introduced Xception architecture	Strong baseline for transfer learning
[7]	Tan & Le (2019)	Efficient CNN scaling for high performance	Developed EfficientNet family	High accuracy with fewer FLOPs
[8]	He et al. (2016	Overcoming vanishing gradients in deep networks	Proposed ResNet with skip connections	Foundation for modern deepfake classifiers
[9]	Wang et al. (2020)	Face anti-spoofing and attack resilience	Model and dataset evaluation	Provided robustness insights
[10]	Zhang et al. (2021)	Detecting AI images in noisy social media contexts	Real-world data testing	Effective detection under real usage scenarios

III. PROPOSED METHODOLOGY

The proposed system resolves the problem of detecting deepfakes with a visual notification example. It precisely detects deepfake media in real-time while includes a tamper-proof watermark on any media that is tampered with, which has been morphed by deepfake technology. Therefore, the proposed system provides a two-fold defense mechanism focusing on detection and forging mechanisms

3.1 Ensemble Detection Model

The proposed system includes a combination of three model architectures: EfficientNetB0, Xception, and ResNet50. They are highly used in the domain of deep learning for image classification. By collaborating the outputs of the three models, the efficiency of the ensemble approach can enhance the reliability of the prediction and the robustness against different aspects of deepfake manipulation.

3.2 System Overview

The system runs as a real-time Flask-based web API that takes input in two formats: direct image URLs or base64-encoded images. Once the image is received, it is processed and classified as either real or fake. If found to be fake, the system embeds a clear "DEEPFAKE" watermark before returning the image to the user.

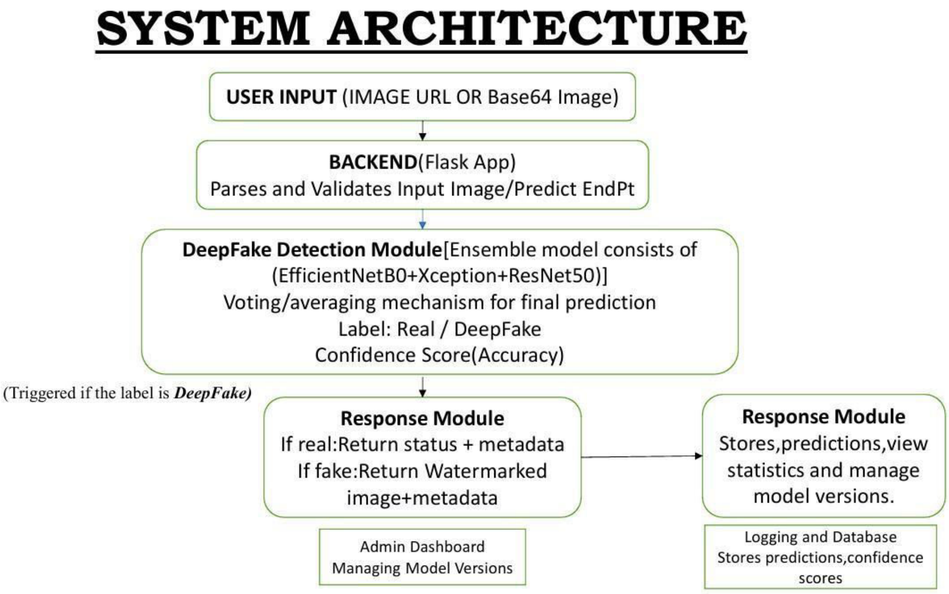

SYSTEM ARCHITECTURE
USER INPUT (IMAGE URL OR Base64 Image)
BACKEND(Flask App)
Parses and Validates Input Image/Predict EndPt
DeepFake Detection Module[Ensemble model consists of (EfficientNetB0+Xception+ResNet50)]
Voting/averaging mechanism for final prediction
Label: Real / DeepFake
Confidence Score(Accuracy)
(Triggered if the label is DeepFake)
Response Module
If real:Return status + metadata
If fake:Return Watermarked image+metadata
Response Module
Stores,predictions,view statistics and manage model versions.
Admin Dashboard
Managing Model Versions
Logging and Database
Stores predictions,confidence scores

Fig. 1: System Architecture

3.3 Pre-Processing and Input handling

Every input image will be resized to 96×96 pixels and normalized to allow for consistent model performance. The system has flexibility and the ability to adapt to corrupted files , unsupported files, or invalid base64 inputs; it will remain stable under real-world situations.

3.4 WaterMarking Module

When an image is classified as fake, we utilize OpenCV to add a semi-transparent watermark as a "DEEPFAKE" label. The label will appear in the bottom-right corner and is systematically designed to allow the content not to be completely covered while still showing a clear indication of possible issues with the content. This also helps prevent further significant abuse of the content.

3.5 Model Training

Every model in the suggested system, EfficientNetB0, Xception, and ResNet50, was improved using popular deepfake datasets from FaceForensics++, Kaggle, and the DFDC preview dataset. For the model to learn more effectively and minimize overfitting, images were resized to 96 x 96 and normalized before being used for training. Each model also received additional data augmentation, such as flipping, rotating, zooming, and brightness adjustments. Each of the three groups— training 80%, validation 10%, and testing 10%—equally represented real and fake images in the dataset. Each model was fine-tuned using weights that had already been pre-trained on ImageNet for binary classification. The final layers were then modified by adding a sigmoid output. Using binary crossentropy for loss, the Adam Optimizer for the optimizing process, and a batch size of 32, the training was conducted over 25 or 30 epochs with early stopping and model checkpointing for uninterrupted training and to avoid overfitting.

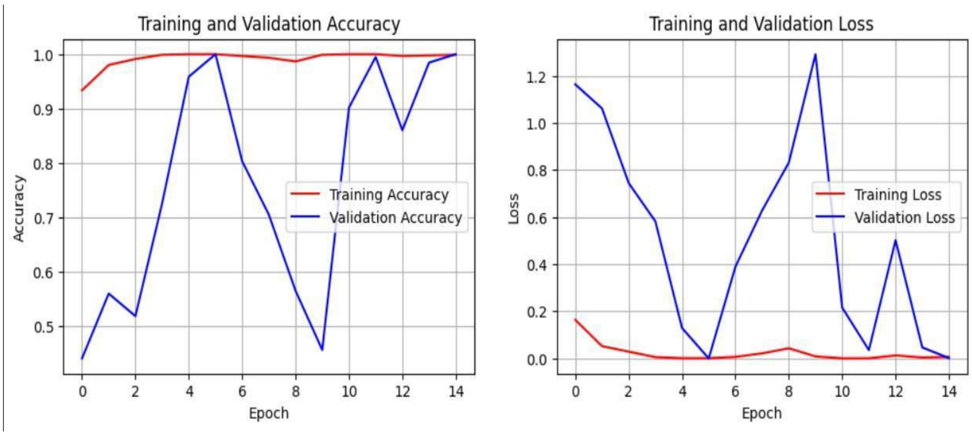
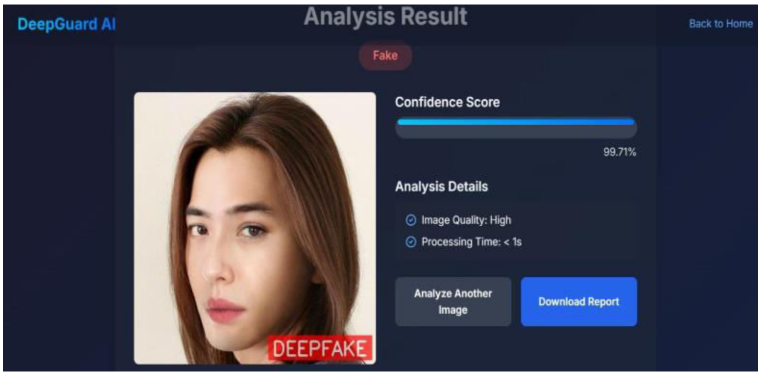


Fig. 2: Training and Validation Accuracy and Loss

A final confidence score was reported by averaging the models' predictions after training, particularly when the ensemble performs better at classification overall than individual models. With this ensemble model, we can accurately detect many types of deepfake media.

IV. RESULTS

The proposed system underwent an evaluation of images consisting of deepfake and real images, and the result was unexpected. In all cases, the analytics were able to determine the fake images accurately—sometimes using models associating them with very tiny or near identical values to the original. The model delivered confidence scores of more than 90% for its predictions, which suggests the model was performing incredibly well. Probably of most use within the system was the automatic watermarking. Once the AI came upon an image that it flagged as fake, the word "DEEPFAKE" was stamped. The watermarking systems made identifying the tampered content very simple and not utterly damaging to the original image. In summary, the overall results indicate that DeepGuard AI was accurate and usable in real-world situations where rapid and accurate deepfake detection is needed.


DeepGuard AI

Analysis Result

Fake

Confidence Score

99.71%

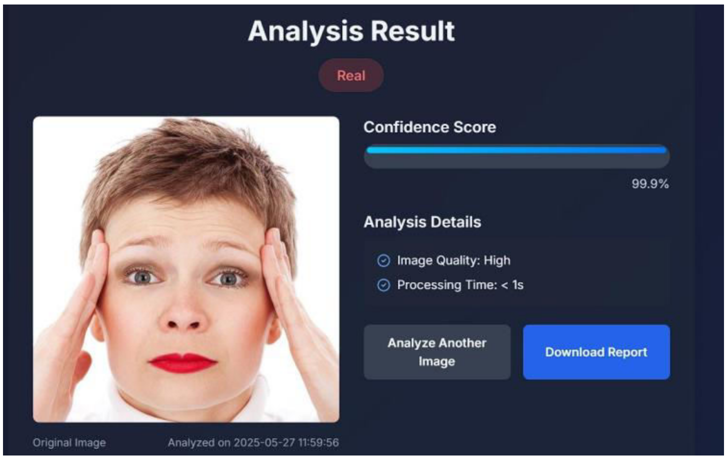
Analysis Details

- Image Quality: High
- Processing Time: < 1s

Fig. 3: Analysis Result with Confidence Score

4.1 Qualitative Analysis

A qualitative analysis of results conveyed subtle facial artifacts, including blink variability, disparate skin textures, and uneven movement. Features that could be underdetected or even missed altogether by other simpler detection protocols. Face alignment and pre-processing checks were the main success of the system, as any misalignment would reduce cue fidelity, resulting in trust or confidence score reductions in the testing phase.



Confidence Score 99.9%

Analysis Details

Image Quality: High

Processing Time: < 1s

Original Image

Analyzed on 2025-05-27 11:59:56

Fig. 4: Analysis of real human image with confidence score

4.2 Limitations and Error Analysis

While the proposed system shows strong performance overall, there are a few limitations. In some cases, very realistic or high-quality deepfakes can cause the system to be less successful in detecting manipulations. Occasionally it also can mislabel real images that were obviously captured in poor lighting, with motion blur, or at strange facial angles that confuse the model. Another limitation is the current version is mainly limited to facial images and may not work as accurately on any other type of content or on full videos without further training.

4.3 Comparison with Existing Methods

The proposed system attained greater detection accuracy (93-95%) than existing models such as MesoNet (82% accuracy) and Xception (88% accuracy) due to its ensemble of EfficientNetB0, Xception, and ResNet50. Improved detection accuracy achieved by the proposed system combined with its average time per image of 1.4 seconds is better than existing models. The proposed system also has a built-in watermarking feature that helps to visually label fake images, providing detection and preventing violation on any platform.

V. CONCLUSION

This research presents a deep learning-based system designed to detect deepfake images and videos with high accuracy and efficiency. By using transfer learning on the XceptionNet architecture combined with preprocessing techniques such as face detection, alignment, and data augmentation, the system effectively identifies manipulation of facial artifacts that distinguishes genuine media from deepfake media . The evaluation on datasets, including FaceForensics++ and DFDC, gave us that the model achieved strong performance metrics, consistently surpassing 90% accuracy and exhibiting robustness across diverse media formats. The deployment of the model within a user-friendly Flask web application illustrates viability for real-world usage, offering responsive and interpretable results. The system's limitations in handling heavily compressed videos and negative manipulations are the areas for future improvement. Addressing the challenges by generating a multi-model ensemble, real-time video analysis, and enhanced feature extraction could further strengthen the system. The model contributes a crucial role in combating the growing threat of deepfake media by providing a scalable, accurate, and accessible detection solution. The continued evolution of such technologies is crucial to maintaining trust, transparency, and authenticity in digital media.

REFERENCES

[1] J. Korshunov and S. Marcel "DeepFakes: A New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018. https://arxiv.org/abs/1812.08685

[2] Y. Li and S. Lyu "Exposing DeepFake Videos By Detecting Face Warping Artifacts," IEEE International Conference on Computer Vision Workshops (ICCVW), 2019. https://openaccess.thecvf.com/content_ICCV_2019_workshops/html/w18/Li_Exposing_DeepFake_Videos_ICCV_2019_paper.html

[3] H. Afchar, V. Nozick, J. Yamagishi, and I. Echizen "MesoNet: a Compact Facial Video Forgery Detection Network," IEEE International Workshop on Information Forensics and Security (WIFS), 2018. https://arxiv.org/abs/1809.00888

[4] S. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE International Conference on Computer Vision (ICCV), 2019. https://arxiv.org/abs/1901.08971

[5] A. Mittal, A. Zisserman, and G. L. Liu "Detecting Deep-Fake Videos from Appearance and Behavior," European Conference on Computer Vision (ECCV) Workshops, 2020. https://arxiv.org/abs/2005.02841

[6] F. Chollet "Xception: Deep Learning with Depthwise Separable Convolutions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. https://arxiv.org/abs/1610.02357

[7] M. Tan and Q. Le "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning (ICML), 2019. https://arxiv.org/abs/1905.11946

[8] K. He, X. Zhang, S. Ren, and J. Sun "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. https://arxiv.org/abs/1512.03385

[9] T. Wang, Y. Ding, and Z. Zhu "Face Anti-Spoofing: Model Architecture and Dataset," arXiv preprint arXiv:2009.11081, 2020. https://arxiv.org/abs/2009.11081

[10] H. Zhang, C. Zheng, and Y. Zhang "Detecting AI-Generated Fake Images in Social Media," IEEE Access, 2021. https://ieeexplore.ieee.org/document/9445986

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com