TruthGuard:DeepFake Face Detection Using

Machine Learning

SWETHA S

*Department of Computer Science*

*and Engineering Panimalar Engineering College*

*Thiruvallur,India*

[ssrswetha2004@gmail.com](mailto:ssrswetha2004@gmail.com)

Dr.SUBEDHA V

*Department of Computer Science*

*and Engineering*

*Panimalar Engineering College Thiruvallur,India*

[subedha@panimalar.ac.in](mailto:subedha@panimalar.ac.in)

SWETHA V

*Department of Computer Science*

*and Engineering Panimalar Engineering College*

*Thiruvallur,India*

[swethavelmurugan06@gmail.com](mailto:swethavelmurugan06@gmail.com)

SHARMILA S

*Department of Computer Science and Engineering*

*Panimalar Engineering College Thiruvallur,India*

[sharmilapanimalar2024@gmail.com](mailto:sharmilapanimalar2024@gmail.com)

Dr.TAMILVIZHI T

*Department of Computer Science*

*and Engineering Panimalar Engineering College*

*Thiruvallur,India*

[tamilvizhi@panimalar.ac.in](mailto:tamilvizhi@panimalar.ac.in)

Dr.JABASHEELA L

*Department of Computer Science and Engineering*

*Panimalar Engineering College*

*Thiruvallur,India*

[ljsheela@gmail.com](mailto:ljsheela@gmail.com)

***Abstract*—Deepfakes are digital-generated images and videos that appear real yet are not. They are generated by sophisticated artificial intelligence methods like deep generative models. Although deepfakes can be employed for artistic purposes like films and entertainment, they also pose a significant threat. They may propagate false information, induce identity theft, money fraud, and diminish o n l i n e media trust a m o n g the public. Conventional detection methods mostly don't work, particularly when the imitated content is ultra- realistic or compact. In this paper, we introduce a deep learning–based framework for detecting deepfakes with the ResNet50 model. The binary cross- entropy loss function is used to train the network to distinguish between images as real or fake. The system is trained and validated using benchmark datasets that include a variety of manipulated faces. The model has a high accuracy rate by detecting small visual features that are visible in forged images. In order to enhance trustworthiness and interpretability, we employ Grad- CAM++, which produces heatmaps indicating the face areas that impacted the decision of the model. This increases the transparency and faithfulness of the system. Our approach demonstrates excellent precision and recalls results and can be applied in use cases like social content verification, digital forensics, and identity Protection.**

**Keywords— Deepfake Detection, CNN, ResNet50, Binary Cross-Entropy, Grad-CAM++, Explainable AI, Digital Media Forensics**

# I.INTRODUCTION

The recent rapid development of artificial intelligence (AI) has resulted in extremely realistic but artificially created digital media. Such deepfake photos and videos, manipulated through sophisticated deep learning algorithms like generative adversarial networks (GANs) and autoencoders,

are produced. Although such technology holds much potential for entertainment, filmmaking, education, and creative industries also have critical risks.

Deepfakes are used to propagate misinformation, engage in identity theft, reputational damage, and shaping public opinion, which results in heavy social, political, and ethical issues. The widespread availability of low-cost computing resources and accessible AI tools has also increased the generation and dissemination of deepfakes on the social media and internet**.**

Conventional statistical and forensic techniques for verifying images tend to fall short when used to deepfakes. This is mainly because modern manipulations are highly detailed, resistant to compression, and visually indistinguishable from real content. Additionally, the gigantic scale on which digital media is disseminated on the web renders manual checking impractical. These issues underscore the imperative for strong, autonomous systems able to identify deepfakes rapidly and effectively.

In this paper, we introduce a deep learning–based system to solve this problem. We use ResNet50, which is a strong convolutional neural network (CNN), as the backbone architecture for real-vs-fake discrimination. Binary cross-entropy is used to train the network loss on benchmark datasets that span a large number of manipulated faces. Our model learns to detect very slight inconsistencies and artifacts within deepfakes that are not visible to the human eye.

To further promote interpretability of the system, we include Grad-CAM++, which produces visual heatmaps indicating the particular facial areas that are responsible for the model's prediction. This not only increases trustworthiness in the system but also gives researchers and users important information about the decision-making process. With the combination of high detection accuracy

with explainability, our method provides a trustworthy solution for deepfake detection with possible uses in digital forensics, identity protection, and content verification online.

# LITERTURE SURVEY

1. The growing commonality of deepfakes in digital media has created serious issues concerning privacy, security, and public trust. Classic deepfake detection methods, even though great at marking forged content, tend to be black-box models that do not output anything about their reasoning process. This makes them less useful in legal and forensic applications where it is just as essential to comprehend the reasoning behind detection as the classification.

Recent studies focus on the implementation of Explainable Artificial Intelligence (XAI) in deepfake detection to enhance interpretability at the cost of little accuracy loss. An example integrates EfficientNetB0 for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal modeling in video sequences[1]. EfficientNetB0 extracts spatial features of video frames, such as textures, colors, and facial patterns, while LSTM networks examine sequential dependency between frames to identify motion or expression inconsistency added through manipulation.

In order to further boost explainability, Grad-CAM (Gradient-weighted Class Activation Mapping) is used to produce heatmaps identifying changed areas of the face, including the eyes and nose. When these heatmaps are summed across frames, the system offers a video-wide view of edited regions, making edit analysis easier and ensuring that detection results are understandable.

The system also includes mechanisms for authentication, e.g., passwords or biometrics, to avoid unauthorized access. Moreover, a web interface enables users to upload video, monitor classification outcome, and engage with visualizations of manipulation patterns. This merging of high accuracy (99.94%) detection and interpretability enhances digital forensics and confidence in AI-driven deepfake detection.

Overall, the literature highlights the significance of hybrid deep learning architectures that capitalize on both temporal and spatial features combined with XAI approaches to develop solid, interpretable, and user-centered deepfake detection systems. These developments are critical for efficient verification of digital content as well as combating misinformation in the media.

1. The advent of deepfake technology has amplified the requirement for effective means of distinguishing artificially doctored images and videos. Deepfakes use Artificial Intelligence(AI) and machine learning, mainly Generative Adversarial Networks (GANs), to generate extremely real but misleading content. GANs include a generator network for generating artificial media and a discriminator network for identifying genuine and fake content, which works throughan iterative learningprocess to

enhance the authenticity of generated media. With advancements in deepfake technology, it becomes more difficult to detect manipulated images and videos, making highly developed detection frameworks a necessity.

Convolutional Neural Networks (CNNs) have been widely used for deepfake detection because they can learn spatial features like textures, edges, colors, and minor visual artifacts that are indicative of manipulation. CNN-based methods examine unusual patterns introduced in the generation phase, allowing the model to distinguish between original and manipulated content. Some dedicated CNN architectures and methods have been developed to improve detection capacity:

Adaptive Manipulation Traces Extraction Network (AMTEN): Proposed by Zhiqing Guo et al., AMTEN is a pre- processing layer that suppresses regular image content and emphasizes manipulation traces. It utilizes adaptive convolution layers to learn manipulation features and optimizes artifact learning in later layers.

Content-Suppressing Convolutional Layers: Bellahassen Bayar et al. introduced convolutional layers that can automatically identify several image manipulations without the need for pre-processing, enhancing robustness and generalization.

Shift-Invariant CNNs: Richard Zhang et al. solved the loss of shift-equivariance of recent deep networks due to common down-sampling layers. Using anti-aliasing filters, these networks achieve consistency regardless of architectures and down-sampling strategies[2].

Recent work shows that CNN-based deepfake detection can generalize to other various datasets and generation methods. For example, a classifier learned on images generated via a single GAN architecture (e.g., ProGAN) has been found to recognize images generated via many unseen architectures, training procedures, and datasets, and is also robust to image scaling, spatial blur, and JPEG compression.

Most traditional methods based on statistical modeling or hand-crafted features cannot identify sophisticated deepfakes. CNN-based methods propose an effective and

scalable solution to deepfake detection by learning intrinsic patterns of manipulation from images. This stimulates the innovation of CNN-based detectors that are minimal in framework but strong in performance. In this paper, a face detection model of deepfakes based on CNN is presented and trained to effectively detect manipulated images, which is beneficial for progress in secure and trustworthy media verification**.**



Fig.1.Datasets

# METHOLOGY

The methodology for detecting deepfake faces proposed here has six significant stages: input dataset, preprocessing, feature extraction, classification, explainability, and result generation.

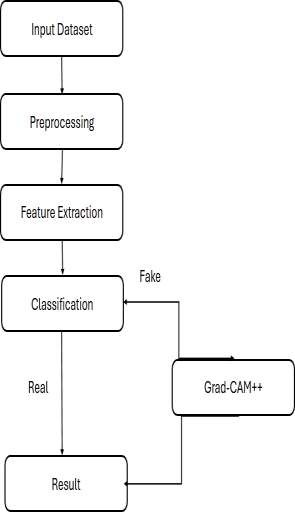


Fig.2.Methodology

* 1. Input Dataset

The first step is to prepare a dataset consisting of real and fake facial pictures. The real pictures are authentic facial samples, while the fake pictures are created based on deepfake algorithms. A balanced dataset is employed so that training and testing can be effective. The dataset is separated into training and validation sets,

which allows the model to generalize well during testing in real-world applications.

* 1. Preprocessing

Preprocessing is done to standardize and eliminate noise before feeding images into the deep learning model. The processes involved are:

Resizing all the images to 224×224 pixels to align with the input size expected by the CNN model.

Normalizingpixel intensity values to the interval [0,1]

for faster convergence during training.

Transformation into tensor format to be easily processed by the model.

Data augmentation methods like rotation, flipping, and scaling are used to increase robustness and decrease overfitting.

* 1. Feature Extraction

Feature extraction is done using a Convolutional Neural Network (CNN). The CNN automatically learns spatial hierarchies of features from the low-level features of edges and textures to the high- level representation of facial structures and inconsistencies. These extracted features play an important role in catching slight artifacts introduced in the generation of fake images.

* 1. Classification

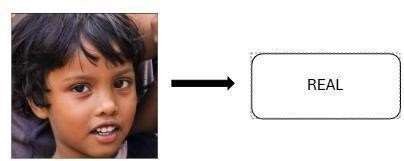
The features extracted are passed through fully connected layers, and classification is performed through a SoftMax activation function. The model gives the probability that the input image is real or fake. The process of decision-making can be summarized below:

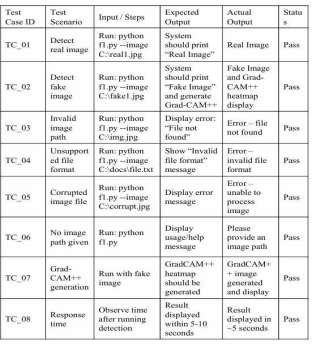
If the image is labeled as Real, then the system directly outputs the result.

If the image is predicted as Fake, its prediction undergoes an additional explainability step by Grad- CAM++.

* 1. Explainability with Grad-CAM++

In order to enhance model explainability and interpretability of predictions, the system includes Grad- CAM++. This method produces a class activation heatmap that identifies the areas of the image most accountable for the prediction. In case of fake images, Grad-CAM++ assists in visualizing manipulated or contradictory areas, producing evidence of tampering. This will increase user trust in the system by ensuring the classification is not a "black box" decision. [3]



*Fig.2.Test Cases*

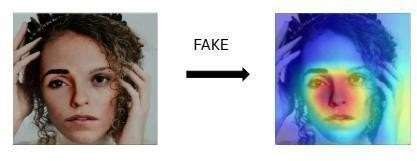
* 1. Result Generation

The last step is to show the classification result. When a path to an image is supplied using the command line interface (CLI):

If the image is Real, then the system outputs

"Real".

If the image is Fake, then the system outputs "Fake" and produces the Grad-CAM++ visualisation. The prediction along with the visual explanation offers both accuracy and explainability in deepfake detection.



*Fig.3.Output*

# RESULT ANALYSIS

The deepfake detection system presented was trained on a dataset of real and fake facial images. ResNet50 was able to learn effectively how to tell the difference between real faces and manipulated ones, registering high accuracy when training and testing. The adoption of binary cross-entropy loss and the Adam optimizer enabled the model to converge effectively, avoiding prediction errors. When evaluating, the system provides a probability score as to whether an image is real or not. Images scoring above a predefined threshold are labeled as real, while those scoring below the threshold are labeled as fake. Experimental results indicate that the model can successfully distinguish fake images even when there are subtle manipulations. For improved interpretability, Grad CAM++ was used for the test images.

The generated heatmaps visually emphasize the facial regions that most contributed to the classification result. These visualizations verify that the model pays attention to relevant features, like eyes, mouth, and other facial contours,which are commonly edited in deepfakes. Overall, the outcomes illustrate that the combination of using ResNet50 for feature extraction and Grad-CAM++ for explainability provides a robust and transparent platform for deepfake detection. The system not only scores well in accuracy but also provides easy-to-understand visual explanations, which are essential in establishing trust in AI- driven detection systems[4].

# CONCLUSION

In this paper, we presented a deep learning-based system to classify real and fake facial images using ResNet50. The model was trained for binary classification using binary cross entropy loss and was optimized through the Adam optimizer, which worked well in distinguishing real images from manipulated images. To achieve more transparency and trust, we employed Grad-CAM++ to determine the regions of the face that influenced the model's predictions. These visual explanations allow users to understand why an image was identified as real or fake and thus make the system accurate and interpretable.

The presented framework offers a complete solution for detecting deepfakes, trading- off credible performance with intelligible visual insights. This approach can be further generalized to video deepfakes and applied in real-world applications for authenticating digital media and security.

# VI .REFERENCE

1. *Vignesh, T., Tarun, P. H., Parthav, R., &*

*Bhargavi,V.(2024).DeepFake machine learning*

*LSTM. Ieee.*

*Face Detection using with*

[*https://doi.org/10.1109/iccsp60870.2024.10544299*](https://doi.org/10.1109/iccsp60870.2024.10544299)

1. *GeeksforGeeks. (2025, July 11). Introduction to Convolution Neural Network.*

*GeeksforGeeks.* [*https://www.geeksforgeeks.org/machine-*](https://www.geeksforgeeks.org/machine-learning/introduction-convolution-neural-network/)[*learning/introduction-convolution-neural-network/*](https://www.geeksforgeeks.org/machine-learning/introduction-convolution-neural-network/)

1. *Vishwakarma, N. (2025, May 1). A guide to*

*Grad- CAM*

*learning.*

*in Analytics*

*Deep Vidhya.*

[*https://www.analyticsvidhya.com/blog/2023/12/grad-cam-*](https://www.analyticsvidhya.com/blog/2023/12/grad-cam-)

[*in-deep-learning/*](https://www.analyticsvidhya.com/blog/2023/12/grad-cam-in-deep-learning/)

1. *GeeksforGeeks. (2025b, July 12). Residual Networks (ResNet) Deep learning. GeeksforGeeks. https:*[*//www.geeksforgeeks.org/deep-learning/residual-*](http://www.geeksforgeeks.org/deep-learning/residual-) *networks- resnet-deep-learning/*