PROJECT PHASE II REPORT

ON

FORTIFYAI : An Explainable Hybrid Deep Learning Framework for Facial Image Forgery Detection

Submitted by

AYSWARYALAKSHMI R(SJC22CS038)

to

the APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

Computer Science and Engineering

Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology, Palai

(Autonomous)

April :: 2026

Declaration

I undersigned hereby declare that the project phase II report on **"FORTIFYAI : An Explainable Hybrid Deep Learning Framework for Facial Image Forgery Detection"**, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala, is a bonafide work done by us under the supervision of **Dr. Joby PP**. This submission represents our ideas in our own words and where ideas or words of others have been included. We have adequately and accurately cited and referenced the original sources. I also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Name and Signature of Student

AYSWARYALAKSHMI R(SJC22CS038)

Place: Choondacherry

Date: 07-04-2026

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY, PALAI

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the report entitled "**FORTIFYAI : An Explainable Hybrid Deep Learning Framework for Facial Image Forgery Detection**" submitted by **AYSWARYALAKSHMI R (SJC22CS038)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering is a bonafide record of the project work carried out by them under my guidance and supervision.

Project Guide
Dr. Joby PP
Professor & Head
Department of CSE

Project Coordinator
Prof. Bino Thomas
Assistant Professor
Department of CSE

Head of Department
Dr. Joby PP
Professor & Head
Department of CSE

Place : Choondacherry
Date : 07-04-2026

Acknowledgement

The success and final outcome of this project phase II required a lot of guidance and assistance from many people, and I am extremely privileged to have received their support throughout the completion of this project. All that we have accomplished is only possible due to their supervision and assistance, and I am sincerely grateful to them.

I would like to express my respect and gratitude to the management of St. Joseph's College of Engineering and Technology for providing us with the opportunity and platform to work on this project.

A special word of thanks goes to our beloved Principal, **Dr. V. P. Devassia**, for providing invaluable support and necessary facilities to carry out this project.

I am also greatly indebted to our Head of the Department and our project guide, **Dr. Joby PP**, Professor in the Department of Computer Science and Engineering, for the invaluable suggestions and encouragement throughout our project work.

I would also like to express our gratitude to our project coordinator, **Prof. Bino Thomas**, Assistant Professor in the Department of Computer Science and Engineering, for his valuable suggestions and guidelines during the entire duration of this project. I truly appreciate his contributions and technical support in preparing this report.

I am thankful and fortunate enough to have received constant encouragement, support, and guidance from all the staff members of the Department of Computer Science and Engineering. Their assistance played a crucial role in the successful completion of our Project Phase II work.

Abstract

The rapid advancement of generative adversarial networks (GANs) and neural face manipulation techniques has led to the widespread creation of highly realistic deepfake images, posing significant challenges to digital media authenticity, identity security, and misinformation control. Existing deepfake detection methods often rely on either deep learning or forensic analysis, resulting in limited generalization across different manipulation techniques. This project presents FortifyAI, a hybrid deepfake detection framework designed to address these limitations by integrating deep learning with multi-signal digital forensic analysis. The system employs an EfficientNet-B3 convolutional neural network enhanced with a perturbation subtraction mechanism and a Fast Fourier Transform (FFT) frequency branch to capture both spatial and spectral artifacts in manipulated images. In addition, a forensic pipeline analyzes multiple evidence signals including error level inconsistencies, sensor noise patterns, PRNU fingerprints, facial landmark geometry, and texture anomalies. These features are combined using an adaptive multi-signal fusion mechanism to generate a final prediction. The system processes input images through face detection, feature extraction, forensic analysis, and decision fusion to classify them as authentic or deepfake. Experimental evaluation demonstrates strong performance across multiple datasets, achieving high accuracy and robust cross-domain generalization. The proposed framework can be applied in domains such as digital forensics, social media monitoring, and cybersecurity. Overall, FortifyAI provides a reliable and scalable solution for detecting deepfake images in real-world scenarios.

Contents

Declaration ii

Acknowledgement iv

Abstract v

List of Figures x

List of Tables xi

List of Abbreviations xii

1 Introduction 1

1.1 Background 1

1.2 Problem Statement 2

1.3 Objective and Scope 2

1.4 Significance of the Study 2

1.5 Proposed Solution 3

1.6 Expected Outcomes 3

2 Literature Review 4

2.1 Physiological Signal-Based Detection 4

2.2 GAN Fingerprint Detection 4

2.3 Mesoscopic Feature-Based Detection 5

2.4 Blending Artifact Detection 5

2.5 Deepfake Benchmark Datasets 5

2.6 Self-Supervised Learning Approaches 5

2.7 Frequency-Based Perturbation Detection 6

2.8 Channel Attention Mechanisms 6

2.9 Deep Neural Network Optimization 6

2.10 Explainable AI Techniques 6
2.11 Optimization Algorithms 7
2.12 Face Detection and Alignment 7
2.13 Efficient Model Scaling 7
2.14 Sensor Noise-Based Forensics 7
2.15 Statistical Steganalysis Methods 7
2.16 3D Geometry-Based Detection 8

3 Methodology **9**
3.1 System Overview 9
3.2 Architecture 10
3.3 Data Flow 11
3.4 Feature Extraction Pipeline 12
3.5 Detection Process 13
3.6 Error Handling and Optimization 14

4 System Study **15**
4.1 Modules 15
4.1.1 Evidence Integrity Module 16
4.1.2 AI Detection Module 16
4.1.3 Error Level Analysis Module 16
4.1.4 Sensor Noise and PRNU Module 16
4.1.5 GAN Upsampling Artifact Module 17
4.1.6 Facial Landmark Module 17
4.1.7 Texture Analysis Module 17
4.1.8 Multi-Signal Fusion Module 17
4.1.9 Explainability Module 18
4.1.10 API Deployment Module 18

5 System Design **19**
5.1 Use Case Diagram 19
5.2 Activity Diagram 21

5.3 Sequence Diagram 22
5.4 ER Diagram 23
5.5 Development Environment 24
5.6 Component Integration 25
5.7 User Interface Implementation 25

6 System Implementation 27
6.1 Development Environment 27
6.2 Component Integration 28
6.3 User Interface Implementation 28
6.4 Database Setup 29
6.5 Testing and Debugging 29
6.6 Deployment 30

7 Experimental Results 31
7.1 Dataset Description 31
7.2 Evaluation Metrics 31
7.3 Quantitative Results 32
7.4 Confusion Matrix 32
7.5 Model Comparison 32
7.6 Cross-Domain Generalization 33
7.7 Real-World Testing 33
7.8 User Interface and System Results 34
7.8.1 Dashboard Interface 34
7.8.2 Admin Panel 34
7.8.3 Detection Result Interface 35
7.8.4 Artifact Visualization 36
7.8.5 Security Module 36
7.9 Explainability Analysis 37
7.10 Discussion 37

Conclusion 37

8 Conclusion **37**

References **38**

ANNEXURE A **40**

ANNEXURE B **47**

ANNEXURE C **56**

List of Figures

3.1 System architecture 11

5.1 Use case diagram of FortifyAI system 20
5.2 Activity diagram of FortifyAI system 22
5.3 Sequence diagram of FortifyAI system 23
5.4 Entity Relationship diagram 24

7.1 User dashboard showing scan statistics and recent analyses 34
7.2 Admin panel displaying system statistics and user feedback 35
7.3 Deepfake detection result with forensic analysis indicators 35
7.4 Grad-CAM and forensic artifact visualization 36
7.5 Two-factor authentication interface 36

List of Tables

7.1 Per-Domain Test Evaluation 32
7.2 Confusion Matrix 32
7.3 Performance Comparison (v4 vs v5) 33
7.4 Cross-Domain Performance 33

List of Abbreviations

API	Application Programming Interface
CNN	Convolutional Neural Network
DL	Deep Learning
FFT	Fast Fourier Transform
FPS	Frames Per Second
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
GUI	Graphical User Interface
MTCNN	Multi-task Cascaded Convolutional Neural Network
PRNU	Photo Response Non-Uniformity
SE	Squeeze-and-Excitation

Chapter 1

Introduction

The rapid advancement of artificial intelligence, particularly Generative Adversarial Networks (GANs), has enabled the creation of highly realistic synthetic media known as deepfakes. While useful in areas like entertainment, these technologies pose serious risks such as misinformation, identity theft, and digital manipulation. This highlights the need for reliable deepfake detection systems. Existing detection methods face limitations in generalization and robustness. Deep learning models perform well on specific datasets but struggle with unseen data, while forensic methods may fail when artifacts are minimal. To address this, this project proposes a hybrid deepfake detection system that combines deep learning and forensic analysis for improved accuracy and reliability.The proposed system, FortifyAI, integrates EfficientNet-based feature extraction with frequency-domain and forensic analysis techniques. Multiple signals such as compression artifacts, noise patterns, and facial inconsistencies are fused to produce accurate predictions suitable for real-world applications.

1.1 Background

Deepfake detection is an emerging field in computer vision and digital forensics. Advances in GANs and face manipulation techniques have made synthetic media highly realistic. Existing approaches include deep learning models for visual feature extraction and forensic techniques for detecting artifacts. However, most systems rely on a single method, limiting effectiveness and highlighting the need for hybrid approaches.

1.2 Problem Statement

Modern deepfake generation techniques have advanced significantly, producing highly realistic images with minimal visible artifacts, making detection increasingly challenging. Existing detection systems often rely on either deep learning or forensic methods, but they suffer from poor generalization and reduced performance on unseen datasets or new manipulation techniques. Additionally, variations in image quality, compression, and noise further impact detection accuracy. This creates a need for a robust and adaptable detection framework capable of handling diverse real-world scenarios while maintaining high reliability.

1.3 Objective and Scope

The objective of this project is to develop a hybrid deepfake detection system that combines deep learning and forensic analysis techniques to improve accuracy and robustness. The system analyzes both visual features and forensic signals, such as compression artifacts, noise patterns, and frequency characteristics, and provides explainable results using techniques like Grad-CAM. The scope of this work focuses on image-based deepfake detection using benchmark datasets such as FaceForensics++ and Celeb-DF. The system is designed to perform effectively under real-world conditions, including variations in image quality and compression. However, video-based detection and real-time processing are beyond the scope of this project and are considered for future work.

1.4 Significance of the Study

This study addresses the growing challenge of ensuring authenticity in digital media. The rapid advancement of deepfake technologies has led to their misuse in spreading misinformation, identity fraud, and cybercrime, making reliable detection systems essential. The proposed FortifyAI system improves detection reliability by combining deep learning with multiple forensic analysis techniques, enabling more accurate and robust identification of manipulated content. Additionally, the integration of explainable AI enhances trans-

parency by providing visual insights into the model's decision-making process, thereby increasing user trust. The system also addresses the limitation of poor generalization in existing methods by incorporating multi-signal analysis. It has wide applications in areas such as cybersecurity, digital forensics, media verification, and social media monitoring, contributing to safer and more trustworthy digital ecosystems.

1.5 Proposed Solution

The FortifyAI system combines EfficientNet-based deep learning with advanced forensic analysis techniques to enhance deepfake detection accuracy. It extracts spatial features from facial structures and frequency-domain features to identify subtle manipulation patterns. In addition, the system analyzes forensic artifacts such as compression inconsistencies, noise patterns, and facial landmark irregularities. These complementary signals are fused using a multi-signal integration approach to produce accurate, robust, and reliable predictions.

1.6 Expected Outcomes

The system is expected to achieve high accuracy across multiple benchmark datasets while demonstrating strong generalization to unseen data. It is designed to provide reliable detection even under real-world conditions such as compression, noise, and varying image quality. The integration of explainable AI techniques enables the system to generate interpretable outputs, improving user trust and transparency in decision-making. Additionally, the system is scalable and can be deployed in practical applications such as cybersecurity, digital forensics, and media verification. Future extensions may include video-based deepfake detection, real-time processing capabilities, and improved adaptability to emerging generative models.

Chapter 2

Literature Review

The Literature Review section provides a comprehensive overview of existing research relevant to deepfake detection and forensic analysis. It establishes the foundation for the proposed FortifyAI system by analyzing prior methods, identifying their limitations, and highlighting the need for a hybrid multi-signal approach. The following subsections discuss key contributions from existing works.

2.1 Physiological Signal-Based Detection

One of the earliest approaches to deepfake detection by analyzing physiological signals, specifically eye blinking patterns. Their method demonstrated that many early deepfake videos lacked natural blinking behavior due to insufficient training data. This approach relies on temporal inconsistencies that are difficult to replicate in synthetic media. While effective for early models, this approach is limited in modern scenarios as newer GAN-based systems have improved significantly and can simulate realistic blinking patterns, reducing the reliability of such temporal features, as proposed by Y. Li et al. [1].

2.2 GAN Fingerprint Detection

The presence of artificial fingerprints left by GAN-generated images was explored in this work. Their study showed that GANs introduce specific frequency-domain artifacts that can be detected through spectral analysis. These fingerprints arise due to the internal architecture and training process of GAN models. This concept forms a strong foundation for frequency-based detection techniques and directly influences the FFT-based forensic analysis used in the FortifyAI system, as proposed by F. Marra et al. [2].

2.3 Mesoscopic Feature-Based Detection

MesoNet, a compact convolutional neural network designed specifically for detecting facial forgeries, focuses on mesoscopic features rather than fine-grained pixel-level details. This enables efficient detection with reduced computational cost and faster inference time. However, its performance is limited when applied to unseen datasets, highlighting the issue of poor cross-domain generalization and reduced robustness in real-world scenarios, as proposed by D. Afchar et al. [3].

2.4 Blending Artifact Detection

This method detects blending boundaries in manipulated images by analyzing pixel-level inconsistencies. It significantly improves detection accuracy by focusing on artifacts introduced during face swapping and image compositing. These inconsistencies often occur at the boundaries of manipulated regions. Despite its effectiveness, the method relies heavily on visible artifacts, which may not always be present in high-quality deepfakes, as proposed by L. Li et al. [4].

2.5 Deepfake Benchmark Datasets

A benchmark dataset for training and evaluating deepfake detection models enabled the development of several high-performing deep learning models. These datasets provide diverse manipulation techniques for robust training. However, models trained solely on such datasets often suffer from overfitting and fail to generalize well to real-world scenarios, especially when exposed to unseen manipulation techniques, as proposed by A. Rossler et al. [5].

2.6 Self-Supervised Learning Approaches

A self-supervised learning approach for deepfake detection reduces dependency on labeled data and improves generalization by learning intrinsic image representations. It leverages

unlabeled data to extract meaningful features. However, it requires complex training strategies and may still struggle with highly realistic GAN-generated images, limiting its effectiveness in certain scenarios, as proposed by Y. Chen et al. [6].

2.7 Frequency-Based Perturbation Detection

This approach utilizes frequency-level perturbations to enhance detection robustness. It demonstrates that frequency-domain features are highly effective in identifying deepfakes, especially under compression and noise conditions. These features remain stable even when spatial artifacts are minimized. This supports the inclusion of FFT-based analysis in modern systems, as proposed by Y. Jeong et al. [7].

2.8 Channel Attention Mechanisms

Channel attention mechanisms enhance model performance by recalibrating feature responses, allowing the network to focus on more informative features while suppressing irrelevant ones. This improves detection accuracy and overall performance, especially in complex datasets with subtle artifacts, as proposed by J. Hu et al. [8].

2.9 Deep Neural Network Optimization

Advanced activation functions and optimization strategies significantly improve deep learning performance and training efficiency. These methods enable deeper architectures to learn complex patterns effectively and form the foundation for modern architectures used in deepfake detection, as proposed by K. He et al. [9].

2.10 Explainable AI Techniques

Visualization techniques such as Grad-CAM provide interpretability by highlighting important regions in an image, helping users understand model decisions. These methods improve transparency and allow validation of whether the model focuses on meaningful features, as proposed by R. R. Selvaraju et al. [10].

2.11 Optimization Algorithms

Improved optimization techniques enhance convergence speed and generalization of deep neural networks. These algorithms reduce overfitting and stabilize training, making them suitable for complex deepfake detection models, as proposed by I. Loshchilov and F. Hutter [11].

2.12 Face Detection and Alignment

Accurate face detection and alignment ensure consistent preprocessing of images, which is critical for reliable deepfake detection. Proper alignment improves feature extraction and model performance across different datasets and variations, as proposed by K. Zhang et al. [12].

2.13 Efficient Model Scaling

Model scaling techniques balance depth, width, and resolution for improved performance and efficiency. EfficientNet provides a strong backbone for feature extraction with high accuracy while maintaining computational efficiency, as proposed by M. Tan and Q. Le [13].

2.14 Sensor Noise-Based Forensics

Sensor noise patterns such as PRNU act as unique fingerprints of real images. These patterns are difficult to replicate in synthetic images, making them highly reliable for forensic analysis. The absence of such patterns in generated images strengthens detection accuracy, as proposed by J. Lukas et al. [14].

2.15 Statistical Steganalysis Methods

Statistical analysis of pixel distributions helps detect hidden anomalies and irregular patterns in images. These techniques are widely used in forensic analysis to identify manip-

ulated content and provide additional evidence for detection systems, as proposed by J. Fridrich et al. [15].

2.16 3D Geometry-Based Detection

Structural inconsistencies in facial geometry can be used to detect manipulated images by analyzing depth and spatial relationships. This approach improves robustness when combined with other detection methods, providing complementary information to pixel and frequency-based techniques, as proposed by H. Li et al. [16].

Chapter 3

Methodology

The methodology chapter outlines the approach used to design and implement the FortifyAI deepfake detection system. This includes a detailed explanation of the system architecture, data flow, and the feature extraction pipeline, which plays a crucial role in identifying manipulation artifacts. Additionally, the detection process explains how the system combines deep learning predictions with digital forensic analysis to classify images as real or deepfake.

3.1 System Overview

FortifyAI is designed as a hybrid and modular deepfake detection system capable of analyzing facial images using both deep learning and digital forensic techniques. The system is structured into multiple modules, each responsible for a specific function such as feature extraction, forensic analysis, and final prediction generation. Key components include:

- **Deep Learning Module**: Uses EfficientNet-B3 to extract spatial features and detect visual manipulation artifacts.
- **Perturbation Analysis Module**: Identifies high-frequency inconsistencies by comparing original and smoothed image representations.
- **Frequency Analysis Module**: Applies Fast Fourier Transform (FFT) to detect spectral patterns introduced by GAN-based image generation.
- **Forensic Analysis Pipeline**: Evaluates multiple evidence signals such as compression artifacts, sensor noise patterns, facial landmark consistency, and texture anomalies.

- **Fusion Module**: Combines outputs from all modules to generate a final classification result.

This modular design ensures flexibility, scalability, and improved detection performance across multiple datasets.

3.2 Architecture

The architecture of FortifyAI, as shown in Figure 3.1, is designed to support efficient deepfake detection by integrating deep learning and digital forensic analysis in a layered manner. The system follows a modular approach where each layer is responsible for a specific stage of processing, ensuring accuracy, scalability, and ease of maintenance.

- **Input Layer**: Accepts the input image uploaded by the user and prepares it for further processing.
- **Preprocessing Layer**: Performs face detection and alignment using MTCNN, followed by image normalization and resizing to ensure consistency in input data.
- **Feature Extraction Layer**: Extracts multiple types of features including spatial features using EfficientNet-B3, residual features using perturbation subtraction, and frequency-domain features using FFT analysis.
- **Forensic Analysis Layer**: Applies multiple forensic techniques such as Error Level Analysis, PRNU-based noise detection, facial landmark consistency checks, and texture anomaly analysis to identify manipulation artifacts.
- **Fusion and Output Layer**: Combines outputs from deep learning and forensic modules using a multi-signal fusion mechanism to generate the final prediction along with a confidence score and visualization.

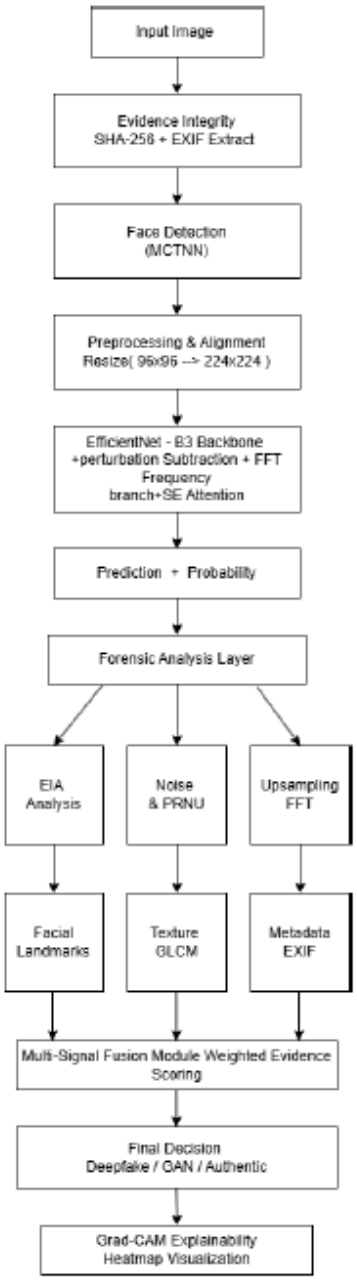


Figure 3.1: System architecture

3.3 Data Flow

The data flow in FortifyAI follows a sequential pipeline from image input to final prediction and output visualization. The process is outlined as follows:

1. The user uploads an image through the system interface for deepfake detection.

2. The preprocessing module detects and aligns the face using MTCNN, followed by resizing and normalization of the image.

3. The feature extraction module processes the image to extract spatial features, residual features, and frequency-domain features.

4. The forensic analysis modules evaluate multiple evidence signals such as compression artifacts, sensor noise patterns, facial landmark consistency, and texture anomalies.

5. The fusion module combines the outputs from the deep learning model and forensic modules to generate a final prediction score.

6. The system produces the output, including the classification result (real or deepfake), confidence score, and visualization such as a heatmap.

This data flow ensures efficient processing and seamless integration between modules, enabling accurate and reliable deepfake detection.

3.4 Feature Extraction Pipeline

The feature extraction pipeline is a core component of FortifyAI, enabling it to identify and analyze manipulation artifacts present in deepfake images. This pipeline consists of several stages:

- **Spatial Feature Extraction**: Uses EfficientNet-B3 to extract high-level visual features such as facial structure, texture consistency, and blending artifacts.

- **Residual Feature Extraction**: Applies perturbation subtraction by comparing the original image with a smoothed version to highlight high-frequency inconsistencies.

- **Frequency Domain Analysis**: Uses Fast Fourier Transform (FFT) to convert the image into the frequency domain and detect spectral patterns introduced by GAN-based image generation.

- **Feature Integration**: Combines all extracted features into a unified representation for further analysis and classification.

- **Residual Feature Extraction**: Applies perturbation subtraction by comparing the original image with a smoothed version to highlight high-frequency inconsistencies.
- **Frequency Domain Analysis**: Uses Fast Fourier Transform (FFT) to convert the image into the frequency domain and detect spectral patterns introduced by GAN-based image generation.
- **Feature Integration**: Combines all extracted features into a unified representation for further analysis and classification.

Each stage in this pipeline contributes to the accurate identification of both visible and hidden manipulation artifacts, improving the overall detection performance.

3.5 Detection Process

The detection process in FortifyAI is responsible for analyzing extracted features and forensic evidence to classify images as real or deepfake. The steps in the detection process include:

1. **Feature Analysis**: The deep learning model processes extracted features and generates an initial probability score.
2. **Forensic Evaluation**: Multiple forensic modules analyze evidence signals such as compression artifacts, noise patterns, and landmark inconsistencies.
3. **Score Normalization**: The outputs from different modules are normalized to ensure consistency across all signals.
4. **Fusion Mechanism**: A multi-signal fusion approach combines the normalized scores to produce a final prediction.
5. **Result Generation**: The system outputs the classification label (real or deepfake), confidence score, and visualization such as a heatmap.

This process enables the system to make accurate and reliable predictions by leveraging both deep learning and forensic analysis.

3.6 Error Handling and Optimization

To ensure reliability and efficiency, FortifyAI incorporates several error handling and optimization strategies:

- **Input Validation**: Ensures that uploaded images are valid and contain detectable facial regions before processing.
- **Fallback Mechanism**: Handles cases where face detection fails by applying alternative preprocessing techniques.
- **Data Augmentation**: Improves model generalization by applying transformations such as rotation, flipping, and brightness adjustment during training.
- **Performance Optimization**: Uses techniques such as mixed precision training and efficient model scaling to reduce computational cost and improve speed.
- **Robustness to Compression**: Ensures that the system performs effectively even on compressed or low-quality images.

These strategies enhance the system's robustness, scalability, and performance across different real-world scenarios.

Chapter 4

System Study

4.1 Modules

The FortifyAI v5 system is composed of multiple interconnected modules designed to perform deepfake detection using a hybrid approach. These modules integrate deep learning techniques with digital forensic analysis to ensure high accuracy, robustness, and adaptability across diverse datasets. Each module performs a specific function and contributes to the overall decision-making process through a structured pipeline.

1. Evidence Integrity Module
2. AI Detection Module
3. Error Level Analysis Module
4. Sensor Noise and PRNU Module
5. GAN Upsampling Artifact Module
6. Facial Landmark Module
7. Texture Analysis Module
8. Multi-Signal Fusion Module
9. Explainability Module
10. API Deployment Module

4.1.1 Evidence Integrity Module

This module ensures the authenticity and integrity of input data before analysis. It computes a SHA-256 hash for each image to uniquely identify and verify the file. Additionally, it extracts EXIF metadata such as camera model, timestamp, and device information. Missing or inconsistent metadata may indicate tampering or synthetic origin. A chain-of-custody mechanism logs all processing stages, ensuring traceability, auditability, and reliability in forensic investigations.

4.1.2 AI Detection Module

This module forms the core of the system by leveraging EfficientNet-B3 for deep feature extraction and classification. It incorporates a perturbation subtraction mechanism to enhance subtle manipulation artifacts and suppress irrelevant noise. A parallel FFT branch captures frequency-domain inconsistencies that are not visible in spatial features. The extracted features are fused and passed through fully connected layers and activation functions to generate a robust prediction score.

4.1.3 Error Level Analysis Module

This module performs Error Level Analysis by recompressing the image and calculating the difference between the original and recompressed versions. Authentic images exhibit uniform error distribution, while manipulated regions show irregular patterns. The module quantifies these inconsistencies using an ELA score, which helps identify localized tampering and editing artifacts, especially in face-swapped images.

4.1.4 Sensor Noise and PRNU Module

This module analyzes intrinsic sensor noise patterns, particularly Photo Response Non-Uniformity (PRNU), which acts as a unique fingerprint of camera sensors. Real images retain consistent PRNU characteristics, while GAN-generated images lack these patterns. The module computes noise residuals and correlation metrics to detect inconsistencies, providing one of the most reliable indicators of image authenticity.

4.1.5 GAN Upsampling Artifact Module

This module focuses on detecting artifacts introduced during GAN-based image generation. By analyzing the frequency spectrum using Fast Fourier Transform (FFT), it identifies periodic patterns and grid-like structures caused by upsampling operations. These artifacts are characteristic of synthetic images and help distinguish them from naturally captured photographs.

4.1.6 Facial Landmark Module

This module evaluates facial geometry using landmark detection algorithms. It extracts key facial points and analyzes symmetry, alignment, and proportional relationships. Deepfake images often exhibit subtle geometric inconsistencies, such as misaligned features or unnatural proportions. These deviations are quantified into a landmark score that contributes to the overall classification.

4.1.7 Texture Analysis Module

This module uses Gray Level Co-occurrence Matrix (GLCM) features to analyze texture patterns in facial regions. It computes statistical measures such as contrast, correlation, homogeneity, and energy. Deepfake images often display unnatural smoothness or repetitive textures, and this module helps capture such anomalies. While not highly discriminative alone, it enhances performance when combined with other modules.

4.1.8 Multi-Signal Fusion Module

This module integrates outputs from all detection modules to produce a final decision. Each signal is normalized and assigned adaptive weights based on its reliability and contribution. Signals such as PRNU and noise patterns are given higher importance due to their robustness. The fusion mechanism reduces false positives and improves cross-domain generalization by combining complementary evidence from multiple sources.

4.1.9 Explainability Module

This module enhances transparency by providing visual explanations of the model's decisions. It uses Grad-CAM to generate heatmaps that highlight regions contributing most to the prediction. These visualizations help users understand whether the model focuses on meaningful features such as facial boundaries and textures. This improves interpretability and builds user trust in the system.

4.1.10 API Deployment Module

This module enables practical deployment of the system through a Django REST API. It provides endpoints for image upload, processing, and result retrieval. The API returns predictions along with confidence scores, forensic analysis outputs, and visualization results. It supports real-time processing, scalability, and easy integration with web and mobile applications, making the system suitable for real-world use.

Chapter 5

System Design

5.1 Use Case Diagram

The use case diagram in Figure 5.1 illustrates the interaction between the User, Admin, and the FortifyAI system, highlighting the key functionalities. The User interacts with the system by logging in, uploading images, tracking detection progress, and providing feedback. The Admin monitors system performance, manages user activities, and reviews feedback, while the System handles internal processes such as preprocessing, deepfake detection, forensic analysis, and result generation.

The workflow begins with the user uploading an image, which is processed by the system using AI-based and forensic techniques. The user can view the results and track progress, while the admin ensures smooth operation by analyzing system performance. This diagram provides a clear view of how different actors interact with the system to enable efficient deepfake detection.

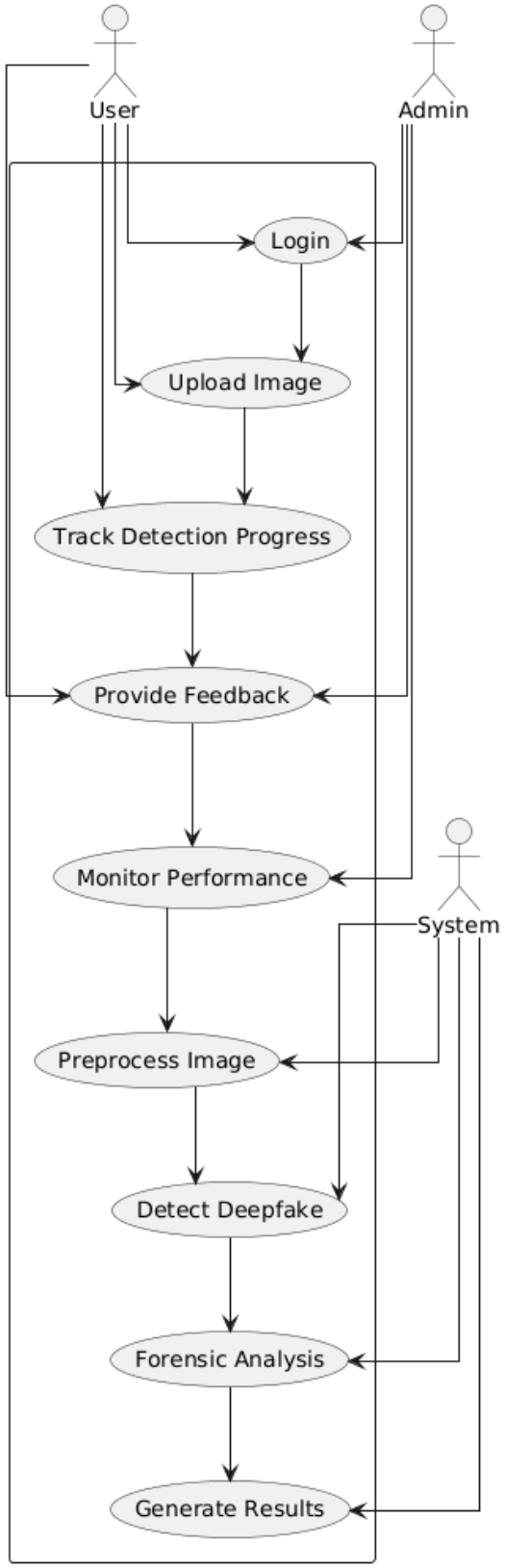


Figure 5.1: Use case diagram of FortifyAI system

5.2 Activity Diagram

The activity diagram in Figure 5.2 illustrates the workflow of the FortifyAI system for detecting deepfake images. The process begins when the user uploads an image through the system interface, followed by validation to ensure the input is suitable for analysis. After validation, preprocessing steps such as face detection and alignment are performed to standardize the image. The processed image is then passed to two parallel modules: the AI detection module and the forensic analysis module. The AI module extracts spatial and frequency features, while the forensic module analyzes artifacts such as compression inconsistencies, noise patterns, and metadata.

The outputs from both modules are combined using a multi-signal fusion mechanism to produce a final decision. Based on this, the system classifies the image as real or deepfake. The result, along with visual explanations, is then displayed to the user. If the input is invalid, the system prompts the user to upload a valid image. This diagram represents both sequential and parallel operations involved in the system, ensuring efficient and reliable deepfake detection.

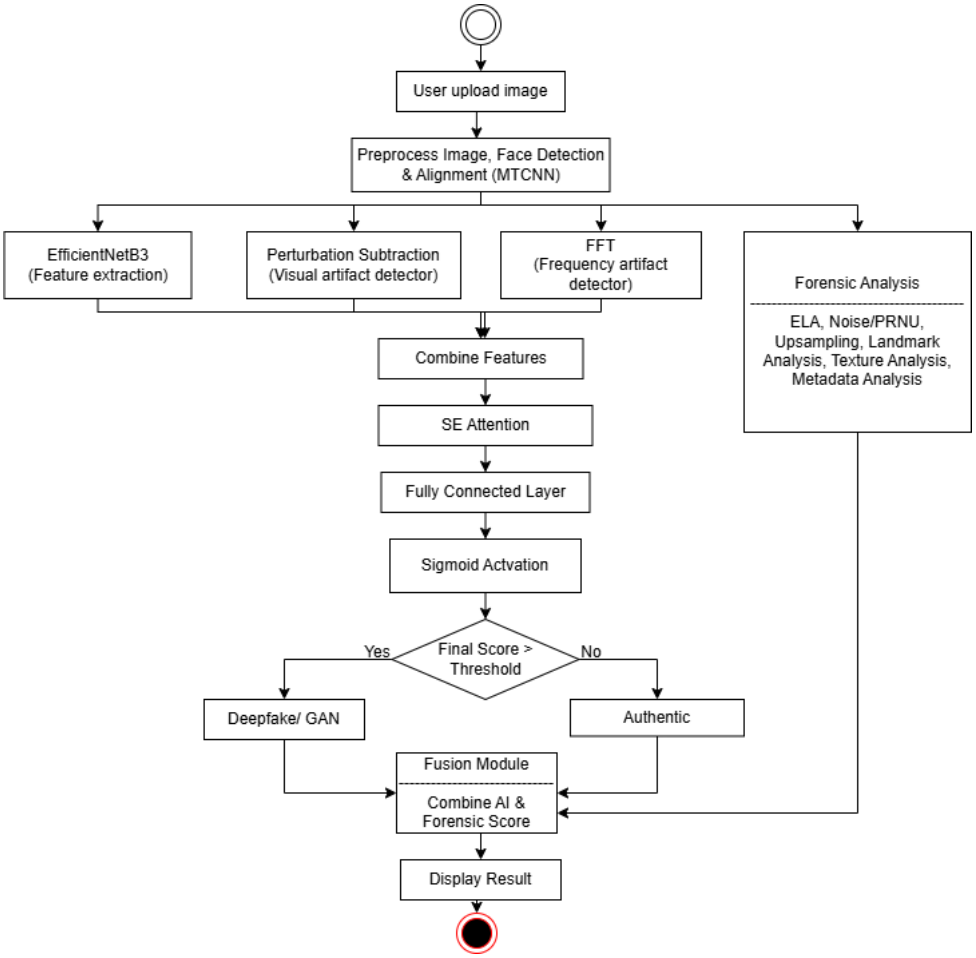


Figure 5.2: Activity diagram of FortifyAI system

5.3 Sequence Diagram

The sequence diagram in Figure 5.3 illustrates the interaction between different components of the FortifyAI system during deepfake detection. The process begins when the user uploads an image through the web application, which is forwarded to the preprocessing module for face detection and alignment.The preprocessed image is then sent to multiple modules in parallel, including the AI detection module (EfficientNet, perturbation subtraction, and FFT) and the forensic analysis module. The AI module extracts deep spatial and frequency features, while the forensic module evaluates artifacts such as ELA, noise/PRNU patterns, upsampling traces, and metadata.

The extracted features are combined and passed through the SE attention mechanism, followed by fully connected layers and sigmoid activation to compute the prediction

score. Based on the threshold, the system classifies the image as authentic or deepfake. The fusion module then combines AI and forensic scores to improve reliability.Finally, the system generates the output, including classification, confidence score, and visual explanation, which is displayed to the user. This diagram highlights both parallel processing and sequential decision-making within the system.

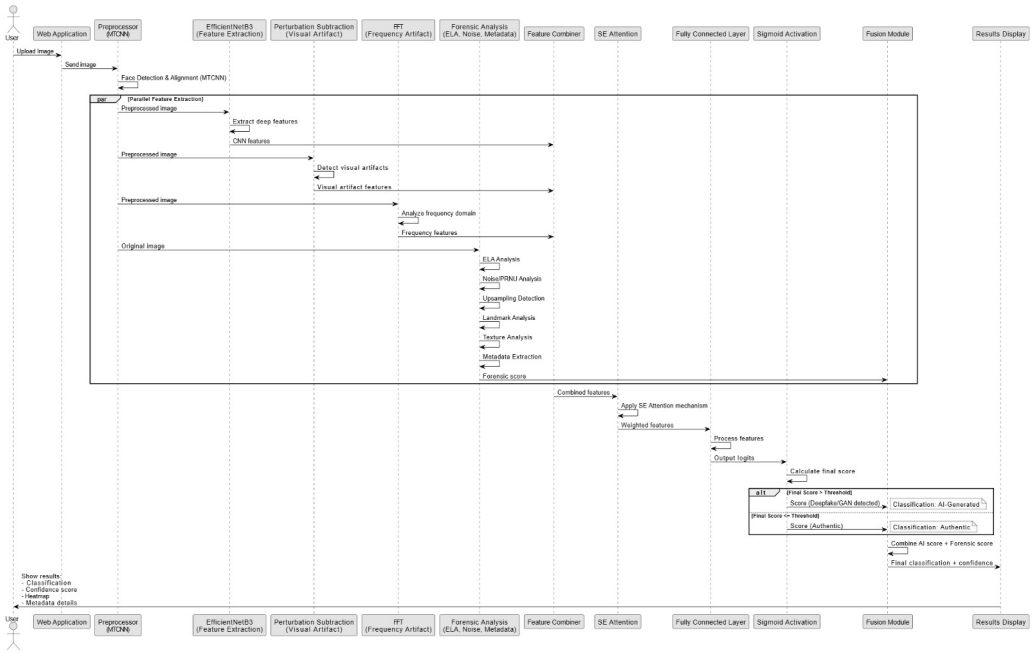


Figure 5.3: Sequence diagram of FortifyAI system

5.4 ER Diagram

The ER diagram in Figure 5.4 represents the data structure of the FortifyAI system and the relationships between different entities. The main entities include User, Uploaded Image, Analysis Result, OTP, and Feedback. The User entity stores details such as name, email, role, and authentication information. A user can upload multiple images, generate OTPs for authentication, and provide feedback. Each uploaded image contains metadata such as filename, size, type, and upload time, and is associated with a single user.

Each image is linked to an analysis result, which includes attributes such as prediction label, confidence score, and various forensic metrics like noise score and compression score. The OTP entity is used for secure authentication, storing details such as code, expiry time, and usage status. The Feedback entity captures user responses, including ratings,

suggestions, and usability metrics. The relationships between these entities illustrate how data flows within the system, ensuring efficient storage, retrieval, and management of user activities and analysis results.

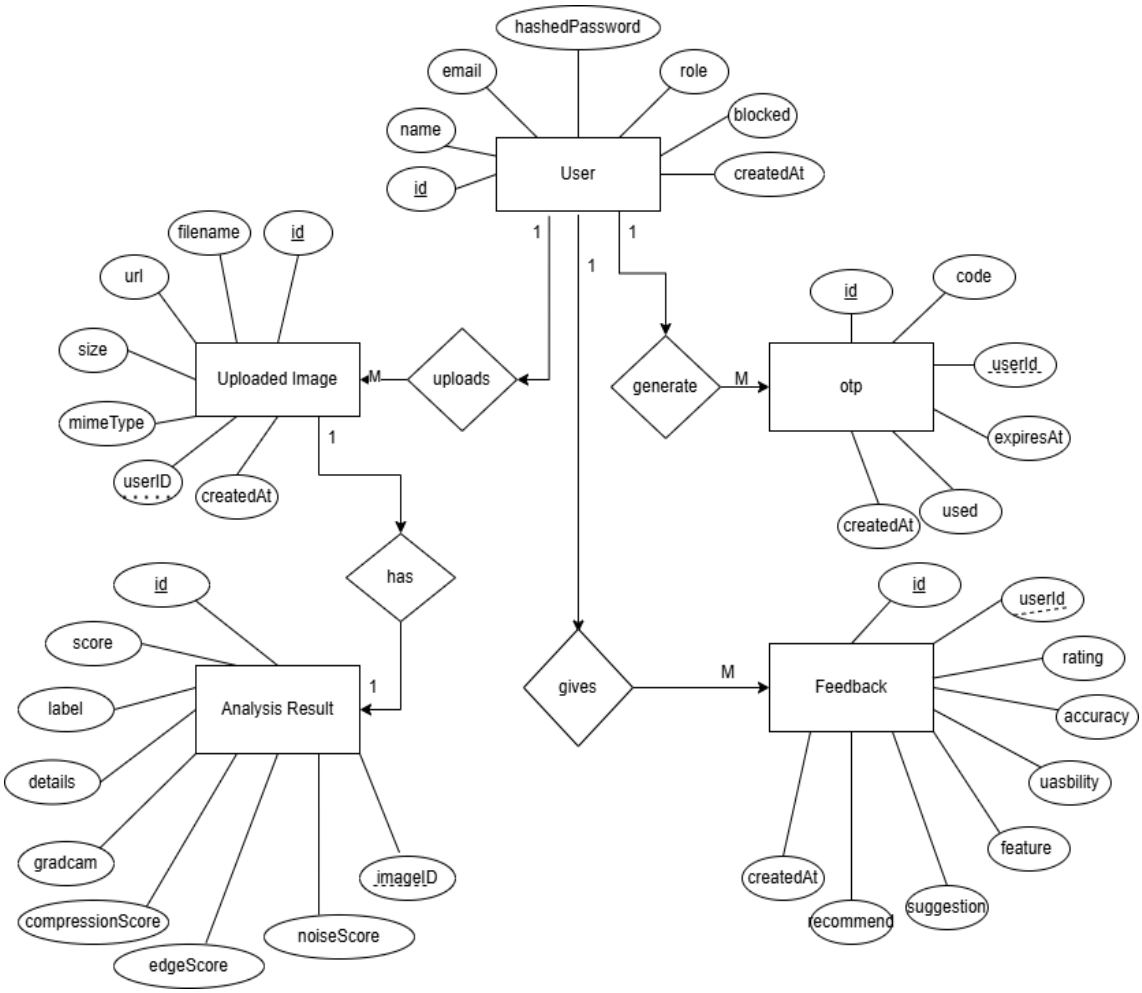


Figure 5.4: Entity Relationship diagram

5.5 Development Environment

The FortifyAI system was developed using modern programming tools and frameworks to ensure efficiency and scalability. The frontend was built using web technologies to provide an interactive interface for image upload and result visualization, while the backend was implemented using Python-based frameworks for integrating deep learning and forensic modules.

Deep learning models were developed using TensorFlow and PyTorch, particularly

EfficientNet architectures. Image preprocessing tasks such as face detection and alignment were performed using OpenCV and related libraries. Forensic analysis modules were implemented to extract features such as ELA, PRNU noise patterns, and FFT-based characteristics.

Version control was managed using Git to track development changes. The system was tested on a standard computing environment with sufficient resources for real-time inference. APIs were used to enable communication between frontend and backend components. Overall, the development environment supported the integration of multiple technologies required for a hybrid deepfake detection system.

5.6 Component Integration

The FortifyAI system follows a modular architecture where multiple components work together to ensure accurate deepfake detection. The process begins with image upload through the user interface, followed by preprocessing steps such as face detection, alignment, and normalization.

The processed image is then passed to two parallel modules: the AI detection module and the forensic analysis module. The AI module extracts spatial and frequency features using deep learning models, while the forensic module analyzes compression artifacts, noise patterns, FFT features, and metadata.

The outputs from both modules are combined in the multi-signal fusion module, which produces a final prediction based on multiple evidence sources. The result generation module then provides classification, confidence score, and visual explanations. This modular integration ensures scalability, efficient processing, and improved detection accuracy.

5.7 User Interface Implementation

The user interface of FortifyAI is designed to be simple and user-friendly, allowing users to upload images and view results easily. The frontend ensures responsiveness across different devices and provides features such as image upload, result display, and feedback options.

After processing, the system displays results including classification (real or deepfake), confidence score, and Grad-CAM visual explanations. The interface also handles errors by notifying users of invalid inputs. Interactive elements such as buttons and progress indicators enhance user experience.

The UI communicates with the backend through APIs, enabling smooth data exchange. Overall, the interface focuses on usability, clarity, and efficient interaction with the system.

Chapter 6

System Implementation

The System Implementation chapter describes how the proposed FortifyAI design is transformed into a fully functional deepfake detection system. This includes details of the development environment, integration of various components, user interface design, database setup, testing procedures, and deployment strategy. The implementation focuses on combining deep learning and digital forensic techniques into a unified and efficient system.

6.1 Development Environment

The development environment for FortifyAI consists of a combination of tools and technologies used for building, training, and deploying the deepfake detection system. The system is primarily developed using Python, which provides extensive support for machine learning and image processing tasks.

Deep learning components are implemented using PyTorch, which enables efficient model training and experimentation. The EfficientNet-B3 model is utilized as the backbone for feature extraction. For face detection and alignment, MTCNN is used due to its accuracy and reliability in detecting facial landmarks.

The backend system is developed using the Django REST Framework, allowing the model to be deployed as an API for real-time predictions. Additional libraries such as NumPy, OpenCV, and Matplotlib are used for image processing and visualization. Version control is managed using Git, ensuring proper tracking of changes and collaborative development.

6.2 Component Integration

The core components of the FortifyAI system are integrated to work together seamlessly, forming a unified detection pipeline:

- **Deep Learning Module**: This component uses EfficientNet-B3 to extract spatial features from facial images and generate an initial prediction score based on learned patterns.
- **Feature Extraction Modules**: These include perturbation subtraction for residual feature extraction and FFT-based frequency analysis for detecting spectral artifacts in images.
- **Forensic Analysis Module**: This module evaluates multiple evidence signals such as compression inconsistencies, sensor noise patterns, PRNU fingerprints, facial landmark geometry, and texture anomalies.
- **Fusion Module**: Combines outputs from the deep learning and forensic modules using a weighted fusion approach to produce the final classification result.

These components are integrated into a single pipeline, ensuring smooth data flow from image input to final prediction.

6.3 User Interface Implementation

The user interface of FortifyAI is designed to provide a simple and intuitive experience for users. The interface allows users to upload images and receive predictions regarding whether the image is real or a deepfake. The UI is developed using web technologies such as HTML, CSS, and JavaScript, ensuring responsiveness and ease of use. The frontend communicates with the backend API to send image data and receive results. The output includes the classification result, confidence score, and visual explanations such as heatmaps generated using Grad-CAM. The design focuses on usability, ensuring that even non-technical users can easily interact with the system.

6.4 Database Setup

A database is implemented to store system-related information such as uploaded images, prediction results, and processing logs. In FortifyAI, **MongoDB**, a NoSQL database, is used due to its flexibility, scalability, and efficient handling of unstructured data such as images and metadata. MongoDB stores data in a **document-oriented format (JSON-like structure)**, which makes it suitable for handling image-related information and dynamic data fields. The database is integrated with the backend using appropriate drivers, enabling seamless communication between the application and storage layer.

The database maintains the following information:

- **Image metadata**: File name, upload time, and image format.
- **Prediction results**: Classification of the image (real/deepfake) along with the confidence score.
- **Forensic analysis outputs**: Optional logging of results from different forensic modules.
- **System logs**: Records of processing history and system operations.

Using MongoDB allows the system to efficiently manage large volumes of data and supports future scalability. It also enables quick retrieval of past results for analysis, debugging, and performance evaluation. Overall, the database plays a crucial role in maintaining system records and supporting the continuous improvement of the model.

6.5 Testing and Debugging

The FortifyAI system undergoes comprehensive testing to ensure reliability and performance. Different levels of testing are performed:

- **Unit Testing**: Individual components such as feature extraction and forensic modules are tested independently.
- **Integration Testing**: Ensures that all modules work together correctly within the detection pipeline.

- **Performance Testing**: Evaluates the accuracy and efficiency of the model using multiple datasets.
- **User Testing**: Validates the usability of the interface and correctness of outputs.

Debugging techniques are applied to identify and resolve issues, ensuring smooth operation of the system under different conditions.

6.6 Deployment

The deployment of FortifyAI involves setting up the system in a production-ready environment where users can access it for real-time deepfake detection. The system is deployed using a backend API built with **Django REST Framework**, allowing integration with web or mobile applications.

The deployment process includes:

- **Model Hosting**: Hosting the trained deep learning model on a server for real-time inference.
- **API Configuration**: Configuring REST API endpoints for image upload and processing.
- **Request Handling**: Ensuring efficient handling of user requests with optimized response time.

This setup enables the system to process images in real-time and deliver accurate predictions. The deployment architecture ensures scalability, allowing the system to handle increasing user demands and support future enhancements.

Chapter 7

Experimental Results

7.1 Dataset Description

The evaluation of the proposed FortifyAI system was conducted using multiple benchmark datasets to ensure robustness and real-world applicability. The datasets include FaceForensics++ (FF++), Celeb-DF v2, and GAN-generated images. The FF++ dataset consists of manipulated videos generated using various deepfake techniques under different compression levels. Celeb-DF v2 contains high-quality deepfake videos with minimal artifacts, making detection more challenging. Additionally, GAN-based datasets include synthetic images generated using models such as StyleGAN.

The preprocessing pipeline involved face detection and alignment using MTCNN. Corrupted or incomplete images were removed, and all images were resized and normalized. The final dataset consisted of approximately 40,000–42,000 images across all domains.

7.2 Evaluation Metrics

The system performance was evaluated using multiple metrics:

- **Accuracy:** Measures overall correctness.
- **Precision:** Indicates reliability of deepfake predictions.
- **Recall:** Measures ability to detect actual deepfakes.
- **F1-score:** Harmonic mean of precision and recall.
- **AUC-ROC:** Measures separability between classes.

7.3 Quantitative Results

The FortifyAI system achieved high performance across all datasets with a mean F1-score of 0.9759 and AUC of 0.9916 according to Table 7.1.

Table 7.1: Per-Domain Test Evaluation

Domain	Accuracy	Precision	Recall	F1	AUC
FF++	93.8%	0.9383	0.9382	0.9382	0.9764
Celeb-DF	98.9%	0.9895	0.9895	0.9895	0.9984
GAN	100%	1.0000	1.0000	1.0000	1.0000
Mean	97.6%	0.9759	0.9759	0.9759	0.9916

7.4 Confusion Matrix

The confusion matrix provides a detailed breakdown of classification performance across different datasets, as shown in Table 7.2.

Table 7.2: Confusion Matrix

Domain	TP	TN	FP	FN
FF++	673	663	49	39
Celeb-DF	706	703	9	6
GAN	712	712	0	0

7.5 Model Comparison

The performance comparison between the previous model (v4) and the proposed model (v5) highlights significant improvements, as presented in Table 7.3.

Table 7.3: Performance Comparison (v4 vs v5)

Domain	v4 F1	v5 F1	v4 AUC	v5 AUC	ΔF1
FF++	0.8408	0.9382	0.9289	0.9764	+0.0974
Celeb-DF	0.9332	0.9895	0.9839	0.9984	+0.0563
GAN	0.8400	1.0000	0.9378	1.0000	+0.1600
Mean	0.8713	0.9759	0.9502	0.9916	+0.1046

7.6 Cross-Domain Generalization

The cross-domain generalization results demonstrate the system's ability to perform effectively across different datasets, as shown in Table 7.4.

Table 7.4: Cross-Domain Performance

Metric	v4	v5
Mean F1	0.8713	0.9759
Mean AUC	0.9502	0.9916
F1 Gap	0.0932	0.0618

The results demonstrate that the system generalizes effectively across datasets and maintains high accuracy even on unseen data.

7.7 Real-World Testing

The system was tested on 11 unseen GAN-generated images, achieving 90.9 percent accuracy (10/11 correctly classified). This confirms strong real-world applicability.

7.8 User Interface and System Results

7.8.1 Dashboard Interface

The dashboard provides an overview of total scans, authentic detections, flagged images, and pending analyses. It allows users to upload images and filter results efficiently shown in Figure 7.1.

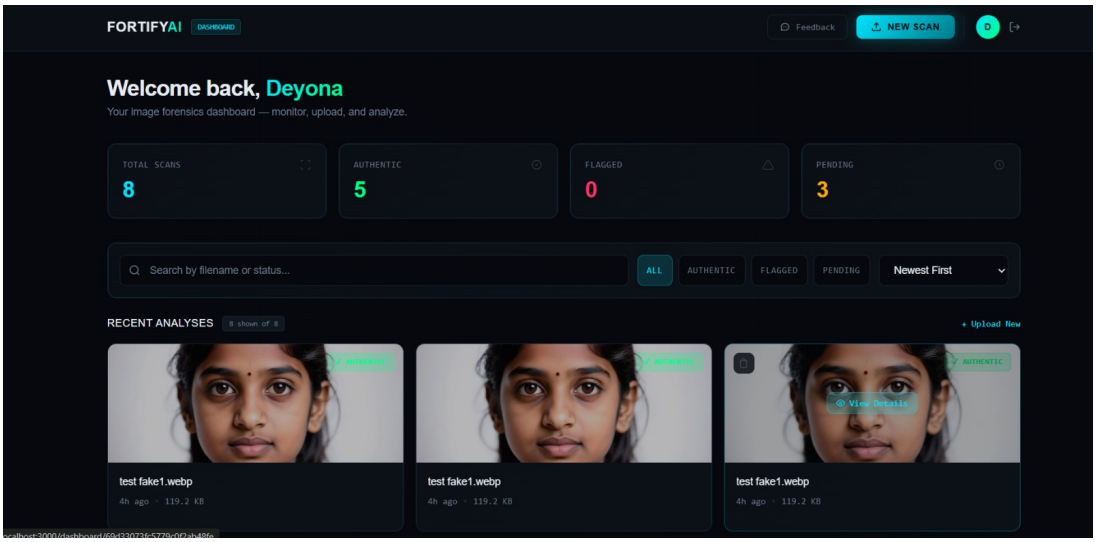


Figure 7.1: User dashboard showing scan statistics and recent analyses

7.8.2 Admin Panel

The admin panel provides system analytics such as total users, uploads, and feedback ratings, helping administrators monitor system performance and evaluate user satisfaction efficiently shown by Figure 7.2.

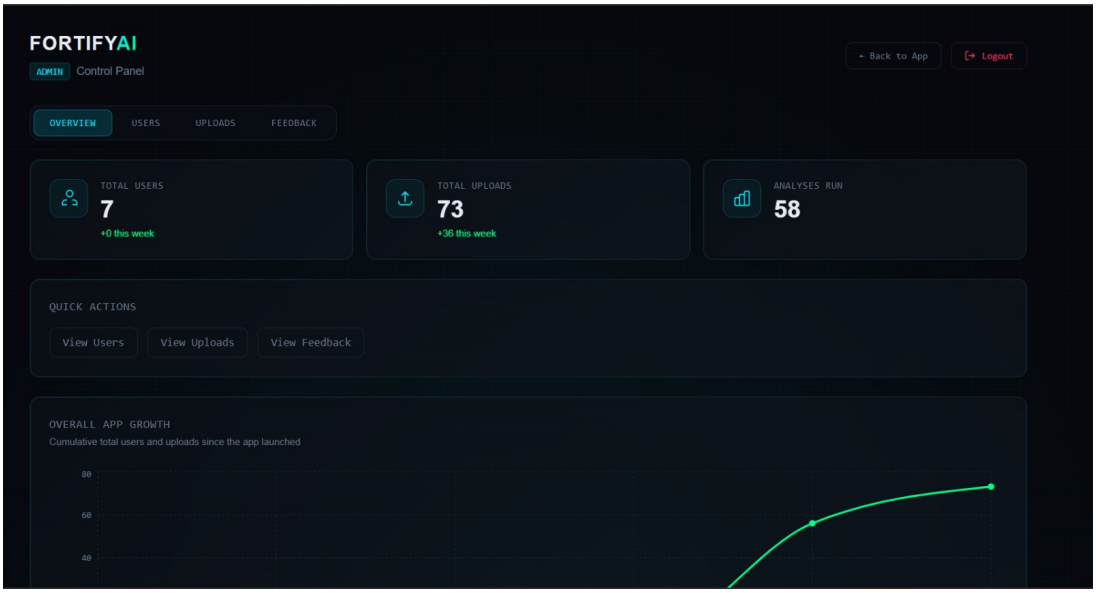


Figure 7.2: Admin panel displaying system statistics and user feedback

7.8.3 Detection Result Interface

The result page shown in Figure 7.3 displays classification and detailed forensic indicators including GAN upsampling, texture analysis, and compression artifacts.

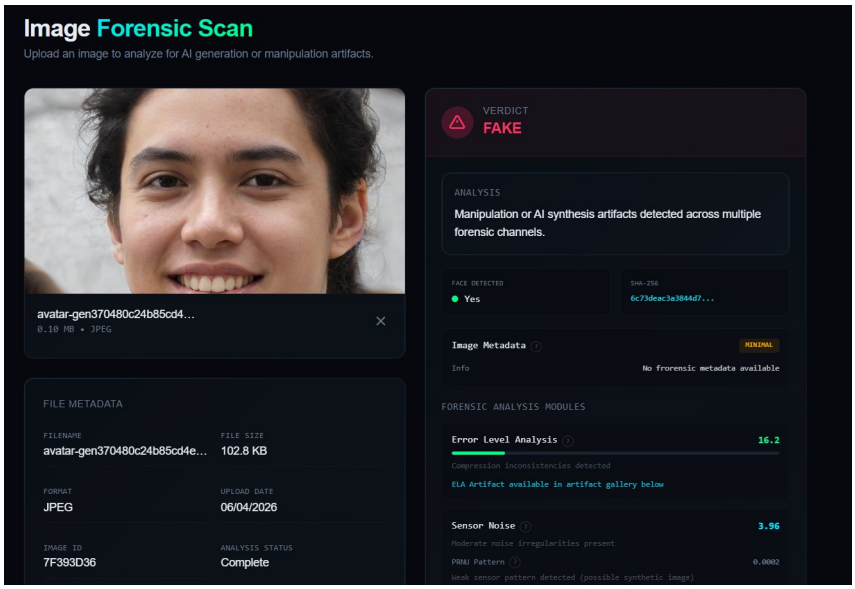


Figure 7.3: Deepfake detection result with forensic analysis indicators

7.8.4 Artifact Visualization

Grad-CAM highlights important regions influencing the model decision shown in Figure 7.4, while ELA and FFT visualizations reveal manipulation artifacts.

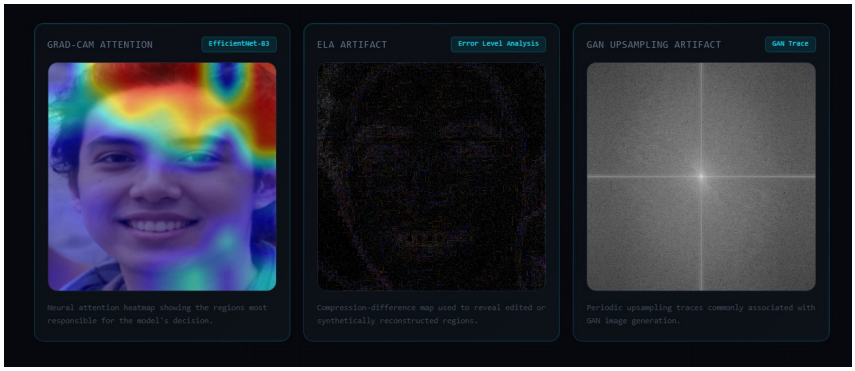


Figure 7.4: Grad-CAM and forensic artifact visualization

7.8.5 Security Module

The system includes OTP-based authentication to enhance security and ensure safe access depicted in Figure 7.5.

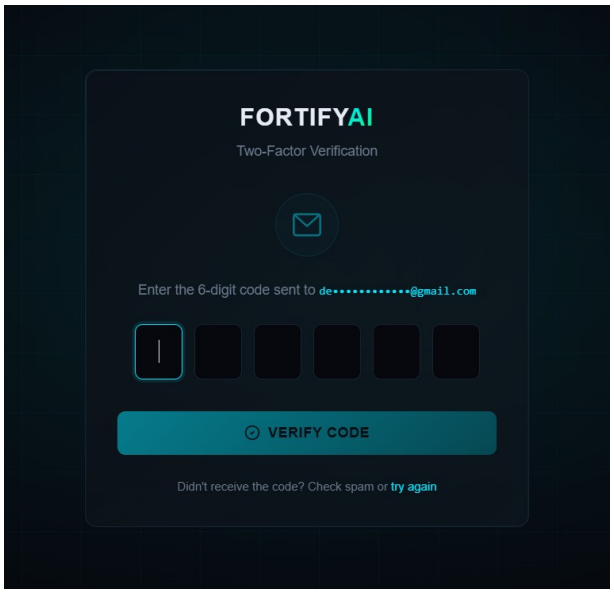


Figure 7.5: Two-factor authentication interface

7.9 Explainability Analysis

The Grad-CAM visualization highlights important facial regions such as the eyes, mouth, and boundaries where manipulation artifacts are most prominent. This helps in understanding which parts of the image contribute most to the model's decision. In addition, forensic visualizations such as Error Level Analysis (ELA) and frequency-domain representations provide further insights into hidden inconsistencies. These explainability techniques improve transparency by allowing users to interpret the model's predictions rather than treating them as black-box outputs. This enhances user trust and makes the system more suitable for real-world applications such as digital forensics and media verification.

7.10 Discussion

The experimental results demonstrate that the proposed hybrid architecture significantly improves deepfake detection performance. The integration of the FFT branch enhances the detection of GAN-generated images by capturing frequency-domain artifacts, while forensic modules contribute to improved robustness by identifying inconsistencies in compression, noise patterns, and facial structures. The combination of deep learning and forensic signals reduces false positives and improves cross-domain generalization, enabling the system to perform effectively across different datasets. The model maintains high accuracy even under challenging conditions such as image compression, noise, and varying resolutions. Overall, the results indicate that the multi-signal fusion approach provides a reliable and scalable solution for deepfake detection, outperforming traditional single-method approaches and making it suitable for practical deployment.

Chapter 8

Conclusion

In this project, a hybrid deepfake detection system, FortifyAI, was developed by integrating deep learning and digital forensic analysis techniques. The system utilizes EfficientNet-based feature extraction along with frequency-domain analysis and multiple forensic signals such as compression artifacts, noise patterns, and facial inconsistencies. The fusion of these complementary features enables accurate and robust detection of manipulated images. Experimental results demonstrate high performance across multiple datasets, with strong generalization and reliability even under real-world conditions such as compression and noise.

Furthermore, the system incorporates explainable AI techniques to provide visual insights into the detection process, improving transparency and user trust. The modular architecture ensures scalability and ease of integration for practical applications in cybersecurity, digital forensics, and media verification. Although the current work focuses on image-based detection, future enhancements may include video-based deepfake detection, real-time processing, and improved adaptability to emerging generative models. Overall, FortifyAI provides an effective and scalable solution to address the growing challenges posed by deepfake technologies.

References

[1] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[2] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?" 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[4] L. Li et al., "Face X-ray for More General Face Forgery Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[6] Y. Chen et al., "Self-supervised Learning for Deepfake Detection," IEEE Transactions on Information Forensics and Security, 2022.

[7] Y. Jeong et al., "FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations," Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[10] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[11] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," International Conference on Learning Representations (ICLR), 2019.

[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning (ICML), 2019.

[14] J. Lukas, J. Fridrich, and M. Goljan, "Digital Camera Identification from Sensor Pattern Noise," IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 205–214, 2006.

[15] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB Steganography in Color and Grayscale Images," IEEE Multimedia, 2001.

[16] H. Li et al., "Face Forgery Detection by 3D Decomposition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

ANNEXURE A

ST. JOSEPH'S
COLLEGE OF ENGINEERING
AND TECHNOLOGY,
- PALAI -
AUTONOMOUS

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CSD416 PROJECT PHASE 2

SECOND REVIEW

FortifyAI: A Lightweight and Generalized Image Forgery Detection System Using Hybrid Deep Learning and Explainable AI

Date : 03/03/2026

Team and Guide Details

Group No:7

Ayswaryalakshmi R
(SJC22CS038)

Deyona Shaji (SJC22CS051)

Namita Jolly (SJC22CS090)

Project Guide:

Dr. Joby P P
Professor

Contents

- Abstract
- Introduction
- Literature Survey
- Existing System
- Proposed System
- System Architecture Diagram
- Methodology
- Hardware Requirement
- Software Requirements
- Functional Requirement
- Non-Functional Requirement
- Design Diagrams
- Module Description
- Partial Output
- Gantt Chart
- Conclusion
- References

Abstract

- Digital image forgery and deepfake manipulation pose serious threats to digital trust and information integrity, demanding robust, generalizable detection systems.
- Existing detectors achieve high accuracy on their training domain but suffer severe degradation under cross-domain conditions, limiting real-world applicability.
- FortifyAI is a hybrid detection framework built on EfficientNet-B3 augmented with perturbation-based subtraction and Squeeze-and-Excitation (SE) channel attention for cross-domain deepfake detection.
- A multi-domain preprocessing pipeline using MTCNN alignment constructs a balanced dataset of 15,838 images from FF++, Celeb-DF, and GAN-synthesised sources with 96×96 identity-suppressing resolution.

Abstract

- FortifyAI achieves a mean accuracy of 87.2%, mean F1-score of 0.8713, and mean AUC-ROC of 0.9502, with Celeb-DF reaching 93.3% accuracy exceeding the 90% target threshold.
- The system is deployed as a Django REST API with Grad-CAM explainability, providing per-prediction spatial heatmaps identifying manipulated facial regions for transparent, real-world deepfake detection.

Introduction

- The rapid advancement of AI has enabled the creation of highly realistic manipulated images, commonly known as **image forgeries or deepfakes**.
- Such forged images can be misused in **social media, identity fraud, misinformation, and digital crimes**.
- Detecting forged facial images has become a **critical challenge** due to increasing image quality and realism.
- This project focuses on **detecting real and fake facial images** using a **deep learning–based approach**.
- A **EfficientNet-B3-based hybrid model with perturbation subtraction and SE attention** is employed for efficient and accurate forgery detection.
- The system also incorporates **Grad-CAM explainability** to visually highlight manipulated regions, improving transparency and trust.

Literature Survey

STUDY	METHODOLOGY	KEY FINDINGS	LIMITATIONS	FUTURE SCOPE
Efficient Deepfake Detection Using AI (2024) Authors: Pradip Kumar Barik, Param Pandya, Saksham Shivansh Gaur--*D*	Used Celeb-DF & UADFV datasets. Extracted video frames → CNN feature extraction → RL agents (PPO, Q-learning, DQN) to optimize augmentations. Applied Explainable AI (GradCAM, ScoreCAM, SmoothGrad).	XceptionNet + PPO agent achieved best performance (AUC ~0.99). Outperformed traditional CNNs in cross-dataset tests. Robust against adversarial attacks.	High computational demand. Dependent on dataset diversity. Limited real-world validation.	Expand datasets to new forgery types. Explore hybrid detection with physiological signals. Focus on real-world deployment for practical use.
Visual Deepfake Detection: Review of Techniques, Tools, Limitations and Future Prospects (2023) Authors: Sourabh Jha, Vijay Kumar--D	Literature review of feature-based, ML-based, and deep learning approaches. Covers detection tools, benchmark datasets (FaceForensics++, Celeb-DF, etc.), and evaluation metrics.	DL (CNN, RNN, Transformer) gives top accuracy; Multi-modal (visual+audio) boosts robustness; Tools: FakeCatcher, MS Video Authenticator.	Detection methods are dataset-specific. High computational cost. Limited generalization to unseen/new deepfake techniques.	Development of lightweight real-time detection systems. Better generalization across datasets and forgery types. Integration of Explainable AI for transparency.
Uncovering DeepFake Images for Identifying Source-images(2024) :Syeda Jannatul Naim1 , Sarker--A	Framework to detect deepfakes + reconstruct source images. Uses VQGAN + CNN + ViT classifier. Dataset: 111k deepfakes + 52k FFHQ images.	Can rebuild original faces used in deepfakes. Better at complex details in base image and simple parts in target image. Low reconstruction error (~0.02–0.03).	Not fully robust (still training). Needs huge data + GPU power. Tested on limited datasets only.	Improve robustness & detail reconstruction. Extend to videos. Build forensic tools for law enforcement. Benchmark with state-of-the-art models.

STUDY	METHODOLOGY	KEY FINDINGS	LIMITATIONS	FUTURE SCOPE
Md Shohel Rana & Andrew H. Sung [2024]: Advanced Deepfake Detection using ML Algorithms--N	Classical ML models like SVM, RF, ERT with features such as HOG and Color Histograms	RF & ERT best (>99% accuracy). ML faster, lighter, more interpretable than DL	Lower accuracy on complex datasets, only visual Deepfakes	Extend to audio/multimodal fakes, real-time detection
Kaiwen Deng & Tingwei Zhang [2025]: FSDNet – Decoupled Forgery Style Network--N	FSDNet with three parts – Contrastive Enhancement, Style Decoupling (CLIP), and Causal Discriminator	91.1% (faces), 84.9% (objects). Outperforms GAN and VGG	Weaker on non-facial images, GPU heavy	Optimize model, expand to more forgery types, real-world use
Shruti Patil, Sujata Joshi, Pratiksha Doke, Shivani Jadhav & Anuja Kulkarni [2024]: Deepfake Detection System Using Deep Learning--N	Implemented CNN, LSTM, and Autoencoder models for image/video deepfake detection	CNN achieved high accuracy for image-based forgeries; Autoencoder useful for anomaly detection	Limited dataset diversity; models weak against unseen manipulations	Larger datasets, hybrid DL models, real-time systems

STUDY	METHODOLOGY	KEY FINDINGS	LIMITATIONS	FUTURE SCOPE
Bambang Sugiantoro [2023]: Deepfake face images: Explainable detection using deep neural networks and class activation mapping--N	Used deep learning models (CNNs/transformers) on benchmark datasets. Preprocessing included normalization and augmentation. Performance measured with accuracy, precision, recall and F1-score	Achieved high detection accuracy and better precision/recall than baseline methods. Models effectively captured visual artifacts and showed robustness across datasets	Limited dataset diversity, high computational cost, reduced performance on sophisticated DeepFakes, focus only on visual cues without multimodal integration	Multimodal detection combining visual and audio, lightweight models for real-time use, larger and more diverse datasets, use of explainable AI for transparency, integration with cybersecurity and legal systems.
Abhishek Kumar Jha et al. [2025]: Deep Learning Based Deepfake Video Detection System--N	Preprocessing (frame extraction, MTCNN, normalization) → CNN (VGG16/ResNet) + BiLSTM → trained on FaceForensics++, Celeb-DF, DFDC	CNN: 85.4%, RNN: 88.1%, CNN-BiLSTM: 92.5% accuracy; ~10 FPS near real-time	Computationally heavy, generalization limits, ethical concerns	Lightweight models; audio–visual multimodal cues; meta-learning; biometric landmark fusion
Deepfake Detection Analysis Based on Video Face Analysis (2024) *Oleh Pitsun, Nazar Melnyk, Khrystyna Lipianina-Honcharenko--A*	Detect facial points (eyes, lips, cheeks). Track patterns (e.g., eye blinks, lip movements). Apply CNNs & RNNs to classify real vs fake.	FaceSwap & Deepfakes Web = strong generators. DFDC, XceptionNet, FakeCatcher = effective detectors. Clues: blink rates, wrinkles, asymmetry.	Newer deepfakes bypass current methods. Needs large data + high computation. Poor generalization to unseen fakes.	Detailed facial analysis (micro-expressions, blood flow). Real-time detection for video platforms. Better datasets & benchmarks. Explore AI + blockchain authentication.

STUDY	METHODOLOGY	KEY FINDINGS	LIMITATIONS	FUTURE SCOPE
Detection of Deepfakes: Protecting Images & Videos(2024): *Ying Tian, Wang Zhou, Amin Ul Haq --- A*	Comparative analysis of CNN & GAN principles. Highlights feature-based, anomaly-based, and deep learning–detection methods. Evaluates multi-modal approaches	CNNs - effective for spotting pixel-level anomalies GAN-based detectors improve accuracy. Multi-modal detection enhances robustness against sophisticated fakes.	GANs need huge datasets and struggle with video consistency, time & resource intensive, current detection methods may fail when exposed to newer, unseen forgery techniques,Arms race problem	Stronger AI models Multi-modal detection Defense strategies
Deepfake Detection Using Deep Learning (2024) – ICACCS.[Focus on deepfake image/video detection using deep learning vs traditional/ML methods]--D	Comparative study of CNN, RNN & Transfer Learning; Datasets: FaceForensics++, FakeCatcher; Preprocessing: cleaning, balancing, feature extraction.	CNNs achieved highest accuracy at pixel-level anomalies; transfer learning helped small datasets; deep learning beats traditional methods.	Computationally expensive, needs large labeled datasets. Focused mainly on deepfakes (faces), poor generalization to unseen forgeries.	Lightweight, real-time models. Extend detection to splicing, copy-move, retouching. Stronger cross-dataset generalization.
Advancements in Deep Learning Techniques for Robust Detection of DeepFakes (2024) Authors: Khushi Shah, Domenec Puig, Kush Mathukiya, Chintan Bhatt --D	Dataset:70k real + 70k fake images; histogram equalization, augmentation; ANN, CNN, VGG16/19, ResNet50, InceptionV3, EfficientNet, U-Net, ViT, MobileNet; Colab, 50 epochs	Transfer learning with MobileNet gave the best accuracy (83%) with lowest loss (0.19). VGG16 also performed strongly (80–87% accuracy). Histogram equalization had minimal effect on results.	Accuracy (83%) still leaves room for improvement. Focus only on deepfake detection, not on other forgery types. Some models (EfficientNet, ResNet50) underperformed. Results dataset-specific, not tested on cross-domain data.	Ensemble methods combining multiple models for better robustness. Incorporate Explainable AI to interpret decisions. Extend to multi-modal detection. Develop more general and real-time detection systems.

STUDY	METHODOLOGY	KEY FINDINGS	LIMITATIONS	FUTURE SCOPE
Deepfake Detection Analysis Based on Video Face Analysis (2024) *Oleh Pitsun, Nazar Melnyk, Khrystyna Lipianina-Honcharenko---A*	Wav2Vec2 (waveform) + UNet (spectrogram). Cross-Modal Fusion (CMFF) + ResNet18 classifier.	Outperforms state-of-art on ASVspoof2021 & In-the-Wild. Fusion captures global + local forgery clues	Needs big data + high compute. Limited dataset testing.	Real-time detection. Larger datasets & better generalization. Lightweight models for deployment.
Exposing Lip-syncing Deepfakes from Mouth Inconsistencies (2024) *Soumyya Kanti Datta, Shan Jia, Siwei Lyu--A*	Extract mouth frames (local + global)→ 3D-CNN + cross-attention → inconsistency loss + classifier.	High accuracy(99%)on FakeAVCeleb; Strong cross-domain (KODF, LSR+W2L); Beats many SOTA lip-sync detectors.	Fails on videos without lip movement or very short clips. Performance drops on face-swap deepfakes.	Add audio modality for better lip-sync detection. Extend to multi-modal detection (audio + video). improve robustness across different deepfake types.
Neha Garg & Rajneesh Rani [2025]: *Recent Advances in Image Forgery Detection: Handcrafted Feature Extraction and Deep Learning-Based Approaches--D*	Handcrafted: SIFT, SURF, DWT, DCT, PCA; ML: SVM, RF; DL: CNNs, GAN detectors, Transformers; Data: CASIA, MICC, CoMoFoD, FF++; metrics used.	Handcrafted works on simple cases but fails on complex ones; DL (CNNs, GAN detectors) improves accuracy; Transformers offer better generalization.	Methods are dataset-specific with poor cross-dataset performance; DL needs large data & high compute; Real-time, lightweight solutions lacking; Few cover all forgery types.	Generalized frameworks for multiple forgeries; Lightweight real-time systems; Cross-dataset robustness & unseen forgery adaptability; Explainable AI for trust.

Dept. of Computer Science & Engineering

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY PALAI AUTONOMOUS

Existing System

- Existing image forgery detection systems depend heavily on complex deep learning models that are computationally expensive and restricted to specific datasets.
- These limitations prevent them from generalizing effectively across various types of image manipulations.
- Moreover, the absence of explainability in their predictions makes it difficult for users to interpret how the system identifies forged content.
- This lack of transparency and generalization reduces their reliability and usability in real-world scenarios.

Dept. of Computer Science & Engineering

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY PALAI AUTONOMOUS

Proposed System

- **Image Forgery Detection Framework:** The proposed system focuses on detecting forged facial images by classifying them as real or fake using deep learning–based feature extraction.
- **Hybrid Architecture with EfficientNet-B3:** EfficientNet-B3 backbone combined with perturbation-based subtraction and Squeeze-and-Excitation channel attention produces a 3,072-d feature vector for binary classification (~13.4M total parameters).
- **Multi-Domain Dataset Construction:** 15,838 face-aligned images from FaceForensics++, Celeb-DF, and StyleGAN-generated sources using MTCNN alignment, stratified frame extraction, and deterministic domain-level balancing.
- **Two-Stage Training Protocol:** Stage 1 (15 epochs, frozen backbone) + Stage 2 (85 epochs, differential LR) with BalancedDomainSampler, MixUp/CutMix augmentation, and AdamW optimizer over 100 epochs.

Dept. of Computer Science & Engineering

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY PALAI AUTONOMOUS

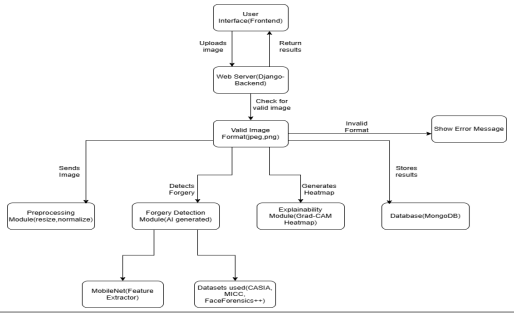
Proposed System

- **Explainable AI Integration:** Grad-CAM is incorporated to visually highlight facial regions that influence the model's predictions, improving transparency and trust.
- **Web-Based Deployment:** Django REST API deployment with Haar Cascade face detection returns binary prediction (AUTHENTIC/DEEPFAKE), confidence score, and base64-encoded Grad-CAM heatmap per inference call.

Dept. of Computer Science & Engineering

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY PALAI AUTONOMOUS

System Architecture Diagram



Methodology

- A labeled dataset containing **15,838 face-aligned images** from FaceForensics++, Celeb-DF, and GAN domains are collected with MTCNN face alignment, stratified frame extraction, and domain-level balancing into 80/10/10 train/val/test splits.
- Images are **preprocessed** at 96×96 pixels to suppress identity memorisation, then upsampled to 128×128 at training time using MTCNN landmark-based face alignment across all three domains.
- **Data augmentation** including horizontal flipping, rotation (+/-10 degrees), JPEG quality degradation, brightness/contrast jitter, MixUp/CutMix (alpha=0.3, p=0.25), and BalancedDomainSampler per batch.
- A **pretrained EfficientNet-B3 backbone with perturbation subtraction** extracts a 3,072-d feature vector; SE attention recalibrates channels; a two-layer MLP head with Dropout(0.4) produces the binary probability via **transfer learning**.
- Custom **classification layers** are added for binary real-vs-fake prediction.

Methodology

- The model is trained using **BCEWithLogitsLoss (label smoothing 0.05, positive class weight 1.2)** and **AdamW (weight decay 3e-4) with OneCycleLR in Stage 1 and CosineAnnealingLR in Stage 2; gradient clipping at max norm 1.0**.
- Performance is evaluated using **accuracy, precision, recall, macro F1-score, and AUC-ROC per domain**.
- **Grad-CAM** is integrated to visualize regions influencing the model's prediction.

Hardware Requirements

Development Phase:

- **Processor:** Intel Core i3 (10th Gen or higher) / AMD Ryzen 3
- **RAM:** 16 GB
- **Storage:** 512 GB SSD
- **GPU:** NVIDIA

Deployment Phase:

- **Processor:** Intel Core i3 Dual-Core or equivalent
- **RAM:** 4 GB
- **Storage:** 50 GB
- **GPU:** Not required

Software Requirements

Development Phase:

- **OS:** Windows 10
- **Language:** Python
- **IDE:** VS Code
- **Libraries:** PyTorch, OpenCV, NumPy, Pandas, Timm, Scikit-learn, Grad-CAM
- **GPU Support:** PyTorch-GPU
- **Dataset:**Kaggle, FaceForensics++, Celeb-DF, Style-GAN

Deployment Phase:

- **Backend:** Python
- **Frontend:**Next.js

Functional Requirements

- **User Authentication:** The system provides **user registration and login** to access the application.
- **Image Upload:** The system allows users to upload **facial images** in supported formats (JPEG/PNG).
- **Image Preprocessing:** Uploaded images are automatically **resized, normalized, and preprocessed** to 96x96 pixels (MTCNN-aligned) and upsampled to 128x128 for EfficientNet-B3 input.
- **Forgery Detection:** The system classifies the uploaded image as **Real or Fake** using a trained deep learning model.
- **Confidence Score Display:** The system displays a **confidence percentage** indicating the model's certainty of the prediction.

Functional Requirements

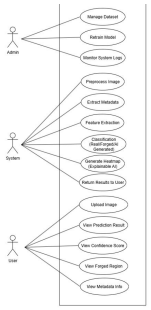
- **Explainability using Grad-CAM:** The system generates a **Grad-CAM heatmap** highlighting image regions that influenced the prediction.
- **Result Visualization:** The prediction result and explanation are displayed through a **simple and interactive web interface**.

Non-Functional Requirements

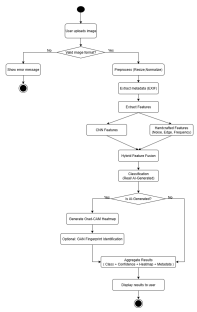
- **Performance:** The system uses a hybrid EfficientNet-B3 model with perturbation subtraction to enable accurate cross-domain image forgery detection with Django REST API deployment.
- **Scalability:** The system can be extended to support larger datasets and additional forgery types in future enhancements.
- **Accuracy & Reliability:** The model is designed to maintain consistent and reliable prediction accuracy across different image samples.
- **Usability:** A simple and intuitive web interface allows users to upload images and view results easily.
- **Security:** Uploaded images are processed temporarily and not stored permanently, ensuring data privacy and preventing leakage.
- **Maintainability:** The codebase is modular, allowing easy updates to the model or addition of new features.
- **Portability:** The application can be deployed on different platforms with minimal modifications.

DESIGN DIAGRAMS:

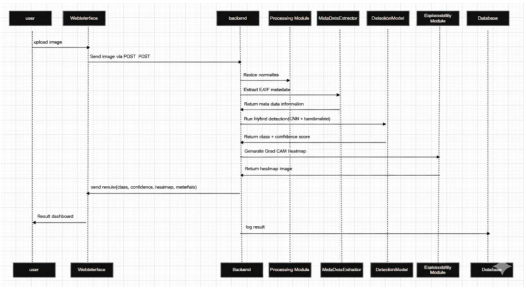
Use Case Diagram:



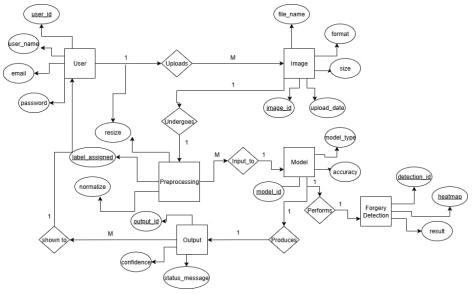
Activity Diagram:



Sequence Diagram:



Entity Relationship Diagram:



Module Description

1. User Authentication Module

- Handles **user registration and login**.
- Ensures only authenticated users can access the forgery detection system.
- Maintains basic user session control for secure access.

2. Image Upload Module

- Allows users to **upload facial images** in supported formats (JPEG/PNG).
- Validates uploaded images before processing.
- Acts as the entry point for forgery analysis.

Module Description

3. Image Preprocessing Module

- Resizes images to **96×96 pixels for preprocessing (identity suppression), upsampled to 128×128 at training**.
- Applies **MTCNN face detection, alignment, and EfficientNet-B3 input preprocessing**.
- Prepares images in the correct format for model inference.

4. Forgery Detection Module

- Uses a **EfficientNet-B3 hybrid model with perturbation subtraction and SE channel attention**.
- Performs **binary classification** to identify images as Real or Fake.
- Generates a **probability score** using sigmoid activation.

Module Description

5. Confidence Estimation Module

- Converts model output into a **confidence percentage**.
- Displays how certain the model is about its prediction.
- Helps users interpret the classification result.

6. Explainability Module (Grad-CAM)

- Generates **Grad-CAM heatmaps** from the last convolutional layer.
- Highlights regions that influenced the model's decision.
- Improves transparency and trust in predictions.

Module Description

7. Result Visualization Module

- Displays prediction result (**Real/Fake**).
- Shows **confidence score** and **Grad-CAM visualization**.
- Provides a simple and interactive user interface.

Evaluation Results

- **FaceForensics++ (FF++):** Accuracy 84.1% | Precision 0.8420 | Recall 0.8409 | F1-Score 0.8408 | AUC-ROC 0.9289 | TP: 287, TN: 268, FP: 62, FN: 43 (660 samples)
- **Celeb-DF (Best Domain):** Accuracy 93.3% | Precision 0.9374 | Recall 0.9333 | F1-Score 0.9332 | AUC-ROC 0.9839 | TP: 324, TN: 292, FP: 38, FN: 6 — exceeds 90% accuracy target
- **GAN-Generated Images:** Accuracy 84.1% | Precision 0.8493 | Recall 0.8409 | F1-Score 0.8400 | AUC-ROC 0.9378 | TP: 303, TN: 252, FP: 78, FN: 27 — highest FP count due to GAN textural overlap
- **Aggregate Cross-Domain Performance:** Mean Accuracy 87.2% | Mean F1-Score 0.8713 | Mean AUC-ROC 0.9502 — assessed as STRONG performance across all three benchmark domains (1,980 total test samples)

Evaluation Results

- **Cross-Domain Generalisation Gap:** F1 Range (Max − Min) = 0.9332 − 0.8400 = 0.0932 — classified as GOOD (<0.15 threshold), confirming acceptable cross-domain stability with only minor domain shift variation
- **Training vs. Test Generalisation:** Best Val F1: 0.8751 | Best Val AUC: 0.9407 (Epoch 99) — Mean Test AUC 0.9502 exceeds training reference, confirming no overfitting and strong generalisation to held-out test distributions

Model Architecture

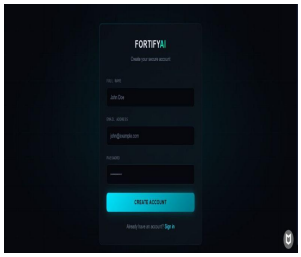
- **EfficientNet-B3 Backbone:** ImageNet-pretrained compound-scaled backbone (~10.7M params) produces a 1,536-d feature vector via global average pooling from 128×128 input. Frozen in Stage 1; final 3 block groups unfrozen in Stage 2.
- **Perturbation Subtraction Mechanism:** f_sub = f_orig − GAP(Backbone(AvgPool(x, k=5))). Exposes high-frequency manipulation residuals. Concatenated with f_orig to produce a 3,072-d spatial fusion vector targeting mid-frequency artifacts.
- **Squeeze-and-Excitation (SE) Attention:** SE block with reduction factor 16 (bottleneck 192) performs channel-wise recalibration on the 3,072-d fusion vector, amplifying manipulation-discriminative channels (~1.18M params).
- **Classification Head:** Two-layer MLP FC(3072→512→1) with Dropout(0.4) and Kaiming normal initialisation; sigmoid output at threshold 0.5 for binary prediction (~1.57M params, total ~13.4M parameters).

Model Architecture

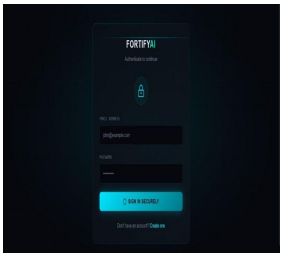
- **Grad-CAM Explainability:** Gradients of sigmoid output w.r.t. EfficientNet-B3 conv_head activations; heatmap bilinearly interpolated to 128×128, JET colormap (alpha=0.4), returned as base64 PNG. Highlights face boundary, eye corner, and mouth regions in deepfakes.
- **Django REST Deployment:** /api/predict/ endpoint accepts multipart image upload; returns JSON with label (AUTHENTIC/DEEPFAKE), confidence probability, natural language explanation, and base64 Grad-CAM PNG. Validated end-to-end via Postman testing.

Output

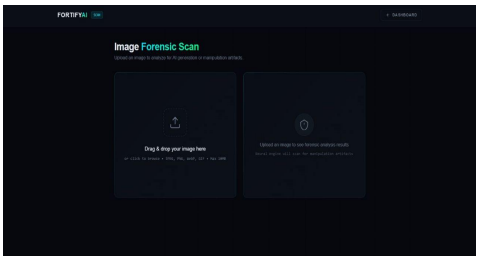
Register Page



Login Page

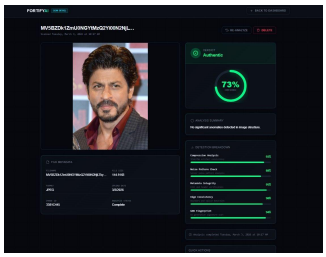


Output

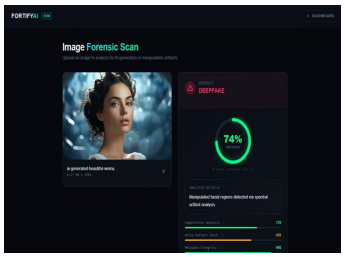


Output

Real Image Detection

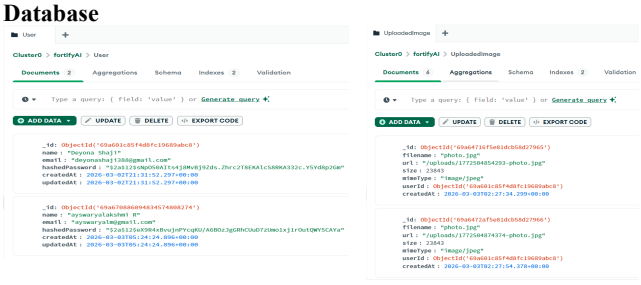


Fake Image Detection



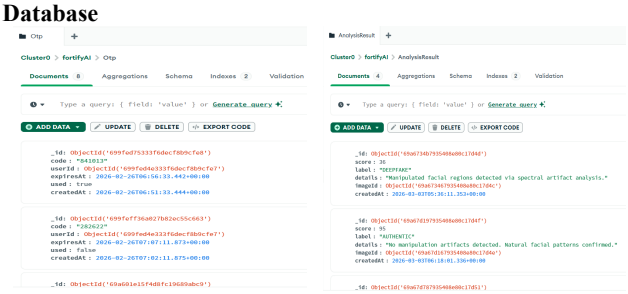
Output

Database

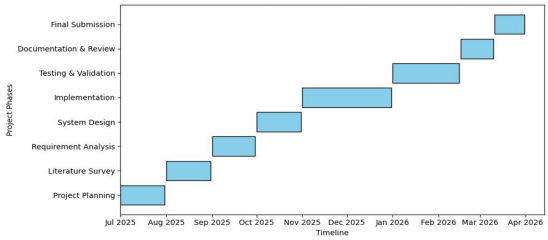


Output

Database



Gantt Chart



Paper Status

- SCAD- Submitted and accepted for presentation
- RAET- Submitted

Conclusion

- FortifyAI successfully demonstrates a cross-domain deepfake detection system using a hybrid EfficientNet-B3 architecture with perturbation-based subtraction and Squeeze-and-Excitation channel attention.
- FortifyAI achieves a mean accuracy of 87.2%, mean F1-score of 0.8713, and mean AUC-ROC of 0.9502 across FF++, Celeb-DF, and GAN domains with cross-domain F1 gap of 0.0932 (GOOD).
- Grad-CAM integration provides per-prediction spatial attribution by activating on EfficientNet-B3's conv_head layer, consistently highlighting facial boundary regions, eye corners, and mouth areas where manipulation artifacts concentrate.

Conclusion

- The Django REST API deployment returns binary predictions with confidence scores and Grad-CAM heatmaps, providing a transparent and interpretable system validated end-to-end via Postman testing.
- Future work includes domain adversarial training, learnable perturbation kernels, frequency-domain augmentation for FF++ compression robustness, and expansion to diffusion-based synthesis and neural radiance field face swapping.

References

[1]Ambika, S., Harini, Y., & Sumanth, B. R. (2025, March). *Real-Time Deepfake Detection using a Hybrid MobileNet-LSTM Model for Image and Video Analysis.* In *8th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 981–986). IEEE. DOI: 10.1109/ICCMC65190.2025.11140856.
[2]Abhishek Kumar Jha et al., "Deep Learning Based Deepfake Video Detection System," 2025.
[3]N. Garg and R. Rani, *"Recent Advances in Image Forgery Detection: Handcrafted Feature Extraction and Deep Learning-Based Approaches,"* 2025.
[4] Kaiwen Deng & Tingwei Zhang, "FSDNet – Decoupled Forgery Style Network," 2025.
[5] H. M. Chen, C. H. Lee, and Y. Tsao, "Wave-Spectrogram Cross-Modal Aggregation for Audio Deepfake Detection," *ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. xx–xx.
[6] *Shruti Patil, Sujata Joshi, Pratiksha Doke, Shivani Jadhav & Anuja Kulkarni [2024]: Deepfake Detection System Using Deep Learning*
[7] M. L. Saini, D. C. Sati, A. Patnaik, R. Kumar, and Mahadev, "Deepfake Detection System Using Deep Neural Networks," *2024 2nd International Conference on Computer, Communication and Control (IC4)*, IEEE, 2024, pp. 1–6.

References

[8] P. K. Barik, P. Pandya, and S. S. Gaur, *"Efficient Deepfake Detection Using AI,"* 2024.
[9]Md Shohel Rana & Andrew H. Sung, "Advanced Deepfake Detection using ML Algorithms," 2024.
[10]S. Jha and V. Kumar, *"Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects,"* 2024.
[11]Y. Tian, W. Zhou and A. U. Haq, "Detection of Deepfakes: Protecting Images and Videos Against Deepfake," *2024 IEEE International Conference on Artificial Intelligence (ICAI)*, 2024, pp. xx–xx.
[12]S. J. Naim and S. T. A. Rumee, "Uncovering DeepFake Images for Identifying Source-images," *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2024, pp. xx–xx.
[13]O. Pitsun, N. Melnyk and K. Lipianina-Honcharenko, "Deepfake Detection Analysis Based on Video Face Analysis," *2024 19th International Conference on Computer Science and Information Technologies (CSIT)*, Ternopil, Ukraine, 2024, pp. xx–xx.

References

[14]S. K. Datta, S. Jia and S. Lyu, "Exposing Lip-syncing Deepfakes from Mouth Inconsistencies," *2024 IEEE International Conference on Multimedia and Expo (ICME)*, Niagara Falls, Canada, 2024, pp. xx–xx.
[15] K. Shah, D. Puig, K. Mathukiya, and C. Bhatt, *"Advancements in Deep Learning Techniques for Robust Detection of DeepFakes,"* 2024.
[16]Bambang Sugiantoro, "Deepfake face images: Explainable detection using deep neural networks and class activation mapping," 2023.

Thank You

ANNEXURE B

Smart Computing, Artificial Intelligence and Data Engineering

An Explainable Hybrid Deep Learning Framework for Facial Image Forgery Detection

Ayswaryalakshmi R
Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai (Affiliated with APJ Abdul Kalam Technological University (KTU))
Kottayam, India
ayswaryalm@gmail.com

Deyona Shaji
Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai (Affiliated with APJ Abdul Kalam Technological University (KTU))
Kottayam, India
deyonashaji388@gmail.com

Namita Jolly
Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai (Affiliated with APJ Abdul Kalam Technological University (KTU))
Kottayam, India
namitajolly333@gmail.com

Dr. Joby PP
Department of Computer Science and Engineering
St. Joseph's College of Engineering and Technology
Palai (Affiliated with APJ Abdul Kalam Technological University (KTU))
Kottayam, India
hodcs@sjcetpalai.ac.in

***Abstract*—The accelerated progress of generative adversarial networks (GANs) and neural face-swapping algorithms has greatly raised the level of realism of synthetic media, also known as deepfakes. Deepfakes are serious threats to digital trust, identity, and information authenticity. This paper introduces FortifyAI, a hybrid deepfake detection system that combines deep learning with a multi-signal digital forensic analysis process. The proposed system uses an EfficientNet-B3 backbone network reinforced with perturbation subtraction and a Fast Fourier Transform (FFT) frequency branch to identify spectral anomalies created during the synthetic image creation process. Besides the deep learning-based detection module, the proposed system also includes a set of digital forensic analysis modules, such as Error Level Analysis (ELA), sensor noise residual analysis, PRNU camera fingerprinting, GAN upsampling artifact analysis, facial landmark consistency analysis, texture anomaly detection, and metadata analysis. A weighted multi-signal fusion strategy is used to combine the results of these modules to produce a reliable final prediction. Experimental results on various datasets show excellent cross-domain generalization and detection performance. The proposed system provides better accuracy and reliability than traditional single-method deepfake detection systems.**

Keywords—Deepfake Detection, Digital Forensics, EfficientNet-B3, FFT Frequency Analysis, Multi-Signal Fusion, GAN Detection, PRNU, Error Level Analysis, Cross-Domain Generalization

I. Introduction

Recent advancements in artificial intelligence have made it possible to produce very realistic synthetic media through the use of generative adversarial networks (GANs) and neural face-swapping algorithms. These technologies can be used to produce fake images or videos that are very similar to real human faces, making it increasingly difficult to distinguish between real and fake media. These synthetic media, also known as deepfakes, pose a significant threat to digital trust and information authenticity.

Deepfakes have been employed in a number of malicious applications, such as political disinformation, identity theft, financial scams, and non-consensual media production. As the technology for the production of deepfakes continues to advance, there is a need for improved detection mechanisms to ensure the integrity of digital media. Early research on deepfake detection involved the identification of physiological anomalies in fake videos. For example, Li et al. suggested the detection of irregular eye blinking, which was commonly absent in early deepfake videos due to a lack of training data [1].

Subsequent works also looked at the analysis of image artifacts and inconsistencies that are left behind by the generative models. Marra et al. showed that images generated by GANs have specific spectral fingerprints that can be used as reliable indicators of synthetic images [2]. However, most of the deep learning-based detection systems are highly dependent on specific features of the dataset, which can cause them to fail on new manipulation techniques.

Broadly, deepfake detection techniques can be divided into two categories: deep learning-based techniques and digital forensic techniques. Deep learning models can learn complex features on large datasets automatically, though they are highly dependent on the ability of the model to generalize over different datasets. Digital forensic techniques, on the other hand, are highly robust, though they may not perform well when the forensic signals are low or are manipulated. To overcome these challenges, this paper presents a hybrid detection system that combines neural network-based

detection with multi-signal forensic analysis. The main concept is that deep learning and forensic signals provide complementary information about manipulated media. By using an adaptive fusion strategy, the proposed system can provide better robustness and generalization performance.

II. RELATED WORK

Several methods have been proposed to detect deepfake images and videos. The initial research was based on the detection of physiological and behavioral inconsistencies of manipulated images and videos. Li et al. proposed a method to detect deepfake videos using unnatural eye blinking patterns that are commonly found in deepfake videos generated using GAN models [1].

Another significant research direction for deepfake detection is the analysis of frequency domain-based artifacts generated using GAN models. Marra et al. showed that GAN models produce unique spectral patterns in the frequency magnitude spectrum of images. These patterns can be identified using frequency analysis techniques, providing significant clues to detect deepfake images.

Deep learning models are also being widely used to detect deepfake images and videos. The MesoNet model, proposed by Afchar et al., uses convolutional neural network models to detect inconsistencies of manipulated images and videos. The Face X-Ray model uses deep learning models to detect blending boundaries of manipulated images. The availability of deepfake image datasets such as FaceForensics++ has enabled researchers to develop powerful deep learning models, such as XceptionNet, to detect manipulated images.

More recent studies have focused on developing methods for improving the generality of the model. Chen et al. have developed a self-supervised learning framework for enhancing the robustness of deepfake detection models for application to new datasets. Jeong et al. have developed perturbation-based deepfake detection methods, which focus on high-frequency manipulation artifacts using frequency-level perturbations.

Apart from deep learning, digital forensic methods have been employed for detecting inconsistencies in manipulated images. Sensor noise analysis and Photo Response Non-Uniformity (PRNU) fingerprinting have been widely used for identifying the source camera for digital images. However, since deepfake images are not captured using actual cameras, they do not exhibit consistent noise patterns.

Although significant progress has been made in developing deep learning-based deepfake detection systems, most existing systems have employed single-source deepfake detection methods. The application of deep learning systems in conjunction with multiple forensic evidence sources has not been explored much. The proposed framework aims to address this gap as integrating both approaches can provide complementary information and significantly improve detection reliability.

III. SYSTEM METHODOLOGY

The proposed system employs a multi-stage methodology that can be used to detect image manipulation by employing deep learning classification along with multiple digital forensic signals. The methodology has five major stages, namely, dataset preparation, preprocessing/face detection, neural feature extraction, forensic signal analysis, and multi-signal fusion scoring. All of these stages provide different information that can be used to make the detection framework more robust.

A. Dataset Preparation and Domain Composition

A critical aspect of deepfake detection is the availability of diverse datasets that represent different manipulation techniques. In this work, three major domains are used to train and evaluate the detection framework:

1. FaceForensics++ Dataset
2. Celeb-DF Dataset
3. GAN-generated Image Dataset

The FaceForensics++ dataset contains manipulated videos generated using several facial manipulation techniques under different compression levels. The Celeb-DF dataset includes high-quality celebrity face-swapping videos that are designed to closely resemble real-world deepfakes. The third domain consists of GAN-generated images paired with authentic facial images to simulate scenarios involving fully synthetic media.

To improve training diversity, multiple frames are extracted from each video sample. Instead of selecting consecutive frames, the system applies stratified temporal frame extraction, where frames are sampled from different segments of the video timeline. This strategy ensures that the dataset captures variations in facial expressions, head poses, lighting conditions, and camera angles. Such diversity helps the model learn robust manipulation features rather than memorizing dataset-specific patterns. The resulting dataset contains tens of thousands of facial images across the three domains, enabling the system to learn generalized representations of both authentic and manipulated media.

TABLE I. DATASET COMPOSITION (V5)

Domain	Type	Videos/Class	Frames/Video	~Total Images
FF++	Video	890	8	~14,240
Celeb-DF	Video	890	8	~14,240
GAN	Image	890	—	~14,240
Total	**—**	**—**	**—**	**~42,720**

B. Face Detection and Alignment

Face localization is a critical step in the deepfake detection process, as the artifacts of manipulation are mainly focused on the facial area. To obtain reliable face crops, the deepfake

detection system employs the Multi-task Cascaded Convolutional Network (MTCNN) for face detection and alignment.

MTCNN is a cascaded three-stage network comprising the following elements:

Proposal Network (P-Net) – responsible for predicting potential face regions.

Refinement Network (R-Net) – refines potential face bounding boxes and filters out incorrect detections.

Output Network (O-Net) – generates final face bounding boxes and corresponding facial landmarks.

This method enables the system to accurately detect facial regions and, at the same time, detect important facial landmarks like eyes, nose, and mouth [12]. Based on the landmarks identified, the system is able to align the face to normalize the orientation of the face. This process ensures that the facial features are aligned in the same manner despite the variance in the images. After the alignment process, the largest face in the image is chosen to avoid multiple faces appearing in the same frame. The aligned face crops are then resized and normalized before being fed into the deep learning model for feature extraction.

C. Image Preprocessing and Resolution Strategy

However, image preprocessing is a significant step in improving model generalization and avoiding identity memorization. In this step, images of faces are initially saved in a low resolution to eliminate high-frequency identity information that may impact the learning process.

Subsequently, images are upsampled to a high resolution prior to training and testing, which is appropriate for convolutional neural network inputs. This process achieves two objectives:

1. It prevents memorization of identity information by the model.

2. It expands the space for detecting manipulation artifacts.

After upscaling, images are processed using normalization techniques to normalize pixel values in the images. This process ensures that changes in brightness, contrast, and color in images do not impact model performance.

D. Neural Feature Extraction

The neural detection part of the proposed system examines the spatial characteristics of the facial image to detect manipulation artifacts. A convolutional neural network is used to extract hierarchical feature representations of the input image.

The network is applied to the aligned facial images and trains to extract discriminative features that separate authentic facial images from manipulated ones. During training, the network is trained to detect subtle visual discrepancies such as:

1. Blending boundaries of manipulated and authentic parts

2. Texture differences in facial skin

3. Distorted facial geometry

4. Artificial patterns created by generative models

Besides spatial characteristics, the system also examines the frequency domain characteristics of the image. Frequency domain analysis is used to help detect periodic artifacts created during image synthesis.

It has been demonstrated in previous research that GAN models create unique frequency patterns as a result of upsampling in generator network architectures [2].

E. Forensic Signal Analysis

In addition to neural detection, the system uses a digital forensic pipeline, which analyzes various image evidence signals. These image signals are used as alternative indicators of image authenticity.

1. Error Level Analysis (ELA)

Error Level Analysis examines the inconsistency in image compression by re-compressing the image and comparing pixel differences between the original image and the re-compressed image. Authentic images have consistent compression errors, while manipulated images have inconsistent errors.

2. Sensor Noise Residual and PRNU Analysis

Digital cameras have unique noise characteristics due to imperfections in their image sensors. This noise pattern, called Photo Response Non-Uniformity (PRNU), helps identify the source camera for an image.

Since GAN-generated images are not captured using digital cameras, they do not have consistent PRNU noise characteristics. The system analyzes the noise residuals and PRNU, which helps identify images that do not have natural noise characteristics.

3. GAN Upsampling Artifact Detection

GAN image generators have an upsampling layer, which uses transposed convolution to increase image resolution. The upsampling process includes periodic artifacts in the image frequency spectrum.

Using the Fast Fourier Transform (FFT) algorithm, the system analyzes the image spectrum to identify periodic peaks, indicating GAN-generated artifacts.

4. Facial Landmark Consistency Analysis

Deepfake generation algorithms may introduce minute irregularities in facial geometry. The proposed system checks the consistency of facial landmark locations to identify geometric inconsistencies between facial landmarks.

5. Texture Anomaly Detection

GAN-generated images tend to have unnatural texture patterns. Statistical texture analysis is employed to identify deviations from natural facial texture distributions.

F. Multi-Signal Fusion Strategy

The final step in this methodology is the fusion of results from the neural detection model and the forensic analysis modules. Each of these modules produces a numerical score that corresponds to the probability of image manipulation.

Since each module focuses on a unique feature of image manipulation, their results are normalized and combined using a weighted fusion technique. This technique uses different weights for each signal depending on their reliability.

For instance, sensor noise and PRNU signals are more discriminative in detecting images generated using GAN, whereas compression artifacts serve as additional evidence. This fusion technique combines all the signals to produce a final score on the probability of a deepfake image, which is used to classify images as either real or fake.

IV. SYSTEM ARCHITECTURE

The FortifyAI framework is developed as a hybrid detection system that combines image classification using deep learning with multiple digital forensic analysis tasks. The framework is organized into two main layers: the AI Detection Layer and the Forensic Analysis Layer, followed by a multi-signal fusion process that produces the final prediction result.

The proposed framework is particularly developed to capture visual manipulation artifacts as well as physical inconsistencies that exist in deepfake images. The combination of neural feature extraction and forensic evidence analysis is expected to enhance the detection robustness of the proposed system.

A. AI Detection Layer

The heart of the neural detection module is founded on the EfficientNet-B3 convolutional neural network, which is commonly employed for image classification tasks owing to its well-rounded scaling of network depth, width, and input resolution [13].

EfficientNet applies a compound scaling approach that enables the network to attain high accuracy with low computational complexity. In the proposed system, the EfficientNet-B3 backbone network is initialized with pretrained weights, enabling the network to benefit from prior knowledge of visual features.

After the input image is processed, the EfficientNet-B3 backbone network generates a high-dimensional feature representation that encodes spatial information about facial structures, textures, and manipulation artifacts.

B. Perturbation Subtraction Mechanism

To further improve the network's capability of identifying subtle manipulation traces, a perturbation subtraction module is incorporated into the network. The module aims to identify and separate high-frequency traces that are introduced during the deepfake manipulation process.

The proposed approach involves generating a smoothed version of the input image by applying an averaging filter to the image. The smoothed image removes the high-frequency details of the image while retaining the facial structure. The proposed network processes the input image and the smoothed image using the same backbone network. The feature maps of the two images are then subtracted to emphasize the differences between the two images.

The above process can be mathematically expressed as:

$$f_{sub} = f_{orig} - f_{pert}$$

where:

- f_{orig} represents the feature vector extracted from the original image
- f_{pert} represents the feature vector extracted from the smoothed image

The resulting residual feature vector emphasizes manipulation artifacts such as blending boundaries and texture inconsistencies. This mechanism improves the model's ability to detect subtle artifacts that may otherwise be overlooked by standard convolutional networks

C. FFT Frequency Branch

Besides the extraction of spatial features, the proposed architecture also includes a Fast Fourier Transform (FFT) frequency branch, which examines the frequency spectrum of the input image.

Frequency analysis has been demonstrated to be effective in identifying GAN-generated images, as generative models tend to introduce periodic patterns during the upsampling process [2]. Such patterns may not be observable in the spatial domain but can be identified when the image is converted to the frequency domain.

The FFT branch carries out the following tasks:

1. Conversion of the input image to grayscale
2. Application of a two-dimensional Fast Fourier Transform
3. Extraction of the log-magnitude spectrum
4. Feature pooling and dimensionality reduction

The extracted frequency features are capable of identifying spectral artifacts introduced by GAN generation processes. The frequency features are then combined with the spatial features extracted by the EfficientNet backbone.

D. Feature Fusion and Attention Mechanism.

Subsequently, after obtaining spatial and frequency domain features, they are concatenated to form a unified feature vector.

In order to enhance the ability of the model to focus on important features, a Squeeze and Excitation (SE) based attention mechanism is employed. This is done by using a Squeeze and Excitation block, which allows for more important features to be given more weight in the classification process.

The Squeeze and Excitation block is composed of two components:

1. Squeeze operation – This involves aggregating spatial information across channels.
2. Excitation operation – This involves learning channel-wise importance weights. By using this weighting, the overall classification is improved.

E. Forensic Analysis Layer

Parallel to this neural detection process, there is a digital forensic analysis pipeline that analyzes various evidence signals for image authenticity.

These forensic modules are:

i. Error Level Analysis (ELA)

ELA detects any inconsistencies in the level of image compression in various areas of the image. Images that have been tampered with often show inconsistent compression compared to genuine images.

ii. Sensor Noise Residual and PRNU

Camera sensors produce unique noise patterns on images. These patterns can be analyzed using PRNU methods to check if the image is genuine or from a real camera sensor [14]. Also, images produced by GAN lack PRNU patterns, which makes this a strong authenticity check for detecting fake media.

iii. GAN Upsampling Artifact Detection

GAN images use upsampling layers in their neural network structure. These layers produce periodic patterns in the frequency spectrum. These patterns can be detected by analyzing the FFT spectrum of the image.

iv. Facial Landmark Consistency

This involves analyzing the consistency in facial landmarks to check for any structural anomalies in the image, which might have been introduced during image tampering.

v. Texture Anomaly Detection

GAN images often show abnormal texture patterns due to their limitations in generating images. These abnormal patterns can be detected by statistical analysis.

F. Multi-Signal Fusion Module

The last step in the architecture involves the fusion of the results from the neural detection model and the forensic modules.

Each module provides a score that corresponds to the likelihood of the image being manipulated. These scores are then normalized to a specific range and fused using a weighted ensemble method.

The fusion score can be computed as:

$$F=w1M+w2E$$

where:

- M represents the neural model prediction
- E represents the combined forensic evidence score
- w1,w2 are weighting coefficients

The final classification decision is based on the fused score.

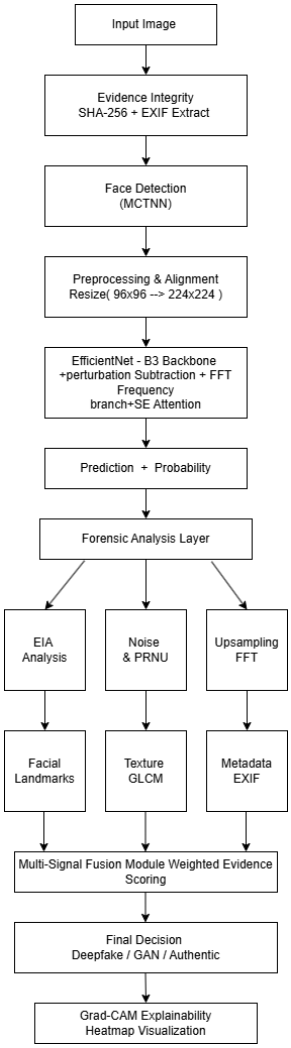


Fig. 1. System Architecture Diagram

V. RESULTS AND DISCUSSIONS

The efficacy of the proposed Fortify AI framework has been tested by employing various benchmark datasets, which are used to assess its ability to detect deepfake images.

1. Face Forensics++
2. Celeb DF
3. GAN-generated image datasets

These are datasets that are comprised of original images, as well as images that are manipulated by various deepfake technologies.

A. Evaluation Metrics

The performance of the system was measured by the following classification metrics:

1. Accuracy
2. Precision
3. Recall
4. F1-score
5. Area Under the ROC Curve (AUC-ROC)

The above metrics will give a complete idea about the detection ability of the proposed model.

B. Experimental Results

The experimental results show that the proposed hybrid framework has excellent detection performance on all datasets.

The model performed:

1. Mean Accuracy: around 97.6%
2. Mean F1-score: 0.9759
3. AUC-ROC: 0.9916

The above results show that the hybrid architecture greatly improves the accuracy of deepfake image detection. The introduction of the FFT frequency branch helped to improve the performance, especially in the detection of GAN-generated images.

TABLE II. V5 PER-DOMAIN TEST EVALUATION (THRESHOLD = 0.5)

Domain	Accuracy	Precision	Recall	F1 (macro)	AUC-ROC
FF++	93.8%	0.9383	0.9382	0.9382	0.9764
Celeb-DF	98.9%	0.9895	0.9895	0.9895	0.9984
GAN	100.0%	1.0000	1.0000	1.0000	1.0000
Mean	**97.6%**	**0.9759**	**0.9759**	**0.9759**	**0.9916**

C. Impact of Frequency Analysis

The FFT component allowed the system to detect the spectral patterns that correspond to the upsampling process performed by GANs. These patterns can still be detected even if pixel-level artifacts are reduced to a certain extent.

The results of the experiment proved that GAN-generated images contain spectral patterns that can be identified in the

frequency domain, as previously suggested by other studies [2].

D. Contribution of Forensic Signals

The analysis modules offered complementary information that aided in the robustness of detection. For instance, the PRNU analysis assisted in identifying images that lack patterns in sensor noise. According to [14], sensor noise patterns are characteristic of camera sensor noise and can thus help in identifying images that have been altered.

The Error Level Analysis assisted in identifying inconsistencies in the compression of manipulated images. The landmark analysis assisted in identifying distortions in the geometry of the images, which are characteristic of deepfake creation.

TABLE III.COMPARISON WITH PUBLISHED DEEPFAKE DETECTORS

System	Architecture	Mean F1	Cross-Domain	Forensic
XceptionNet [5]	Xception	~0.81 *	Limited	No
F3Net	ResNet + Freq	~0.88 *	Moderate	No
RECCE	Reconstruction	~0.87 *	Moderate	No
FortifyAI v4	EB3 + Perturbation	0.8713	Good	No
Proposed system v5	**EB3 + FFT + Forensics**	**0.9759**	**Excellent**	**Yes**

VI. CONCLUSION

The rapid progress in generative models has significantly improved the quality of synthesized media, thereby making the detection of deepfakes a critical problem in modern digital media forensic analysis. Generative adversarial networks and neural face-swap methods have shown promising results in generating realistic synthesized media, which is difficult to distinguish from real media using conventional detection methods. With the increasing availability of deepfake technology, the necessity for robust and reliable detection systems is on the rise. In this paper, a hybrid framework for detecting deepfakes has been proposed to overcome the limitations of conventional methods for detecting deepfakes. The proposed framework combines deep learning-based image classification with various digital forensic analysis methods for reliable detection. The framework is based on a convolutional neural network (CNN) architecture named EfficientNet-B3, which is known for providing efficient feature extraction for image classification problems with high accuracy [13]. The proposed architecture is enhanced by incorporating a perturbation subtraction mechanism and a frequency analysis branch, which allows the CNN to learn from spatial and frequency domain information introduced during GAN-based image generation.

Another strength of the proposed framework is the use of multiple forensic signals in conjunction with neural detection. In the proposed framework, the analysis pipeline for digital forensics considers various evidence types, including Error Level Analysis, sensor noise residual patterns, PRNU-based camera fingerprints, GAN upsampling artifacts, facial landmark consistency, and texture anomalies. These forensic signals are used to identify inconsistencies in the manipulated images. For instance, PRNU-based analysis has been used extensively in digital forensics to identify noise patterns in images taken by cameras and verify the authenticity of the images. Since synthetic images are not captured by cameras, they do not have noise patterns, and hence, PRNU-based analysis can be used to identify manipulated images. The experimental evaluation of the proposed framework has demonstrated that the proposed framework has excellent detection capabilities. The use of frequency analysis has greatly enhanced the detection capabilities of the proposed framework in identifying GAN-based synthetic media. As discussed in the literature, the frequency analysis of synthetic media has shown that the proposed framework can effectively identify synthetic media, and the experimental evaluation has validated the literature that frequency analysis can be used to identify synthetic media. Another significant contribution of this research is the multi-signal fusion mechanism, which is used to combine the outputs of neural detection and forensic modules to make a final classification decision. The system is less prone to false predictions that may arise when a single detection method is used. This is because the system uses a combination of multiple evidence sources.

Although the results obtained by the proposed system are promising, there are still challenges that need to be addressed. As the generative models continue to advance, the future deepfake generation models will be less detectable, making the detection process even more challenging. For instance, diffusion-based image synthesis models are being developed as a new alternative to GANs. Future work, therefore, includes extending the proposed framework to include detection of diffusion-based synthetic media, improving robustness against adversarial attacks, and improving the fusion strategy using techniques such as adaptive weighting and auto-calibration. Moreover, extending the proposed framework to include real-time deepfake detection in videos and using techniques such as visual attribution maps in explainable AI could further help in improving the usability of the proposed framework in digital forensic investigations.

In conclusion, it is evident that the proposed framework shows great promise in using deep learning and digital forensic techniques to provide a powerful solution in deepfake detection. The proposed framework uses multiple sources of information and deep learning techniques to provide a robust solution in deepfake detection and could be used in real-world applications to improve the reliability and robustness of deepfake detection techniques.

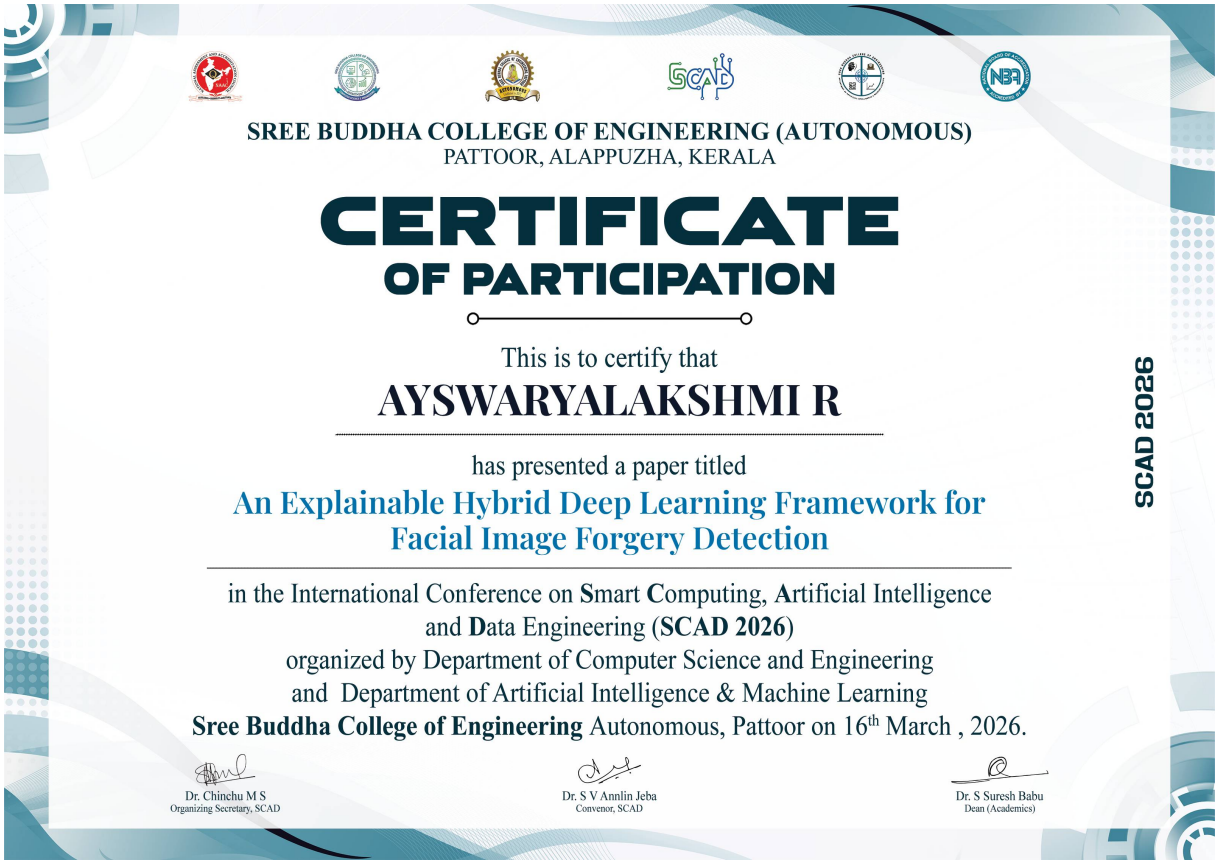
REFERENCES

[1] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," IEEE WIFS, 2018.
[2] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?" IEEE MIPR, 2019.
[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE WIFS, 2018.
[4] L. Li et al., "Face X-ray for More General Face Forgery Detection," CVPR, 2020.
[5] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV, 2019.
[6] Y. Chen et al., "Self-supervised Learning for Deepfake Detection," IEEE TIFS, 2022.
[7] Y. Jeong et al., "FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations," AAAI, 2022.
[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," IEEE TPAMI, 2020.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers," ICCV, 2015.
[10] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," ICCV, 2017.
[11] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularisation," ICLR, 2019.
[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE SPL, vol. 23, no. 10, pp. 1499–1503, 2016.
[13] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," ICML, 2019.
[14] J. Lukas, J. Fridrich, and M. Goljan, "Digital Camera Identification from Sensor Pattern Noise," IEEE TIFS, vol. 1, no. 2, pp. 205–214, 2006.
[15] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB Steganography in Color and Grayscale Images," IEEE Multimedia, 2001.
[16] H. Li et al., "Face Forgery Detection by 3D Decomposition," CVPR, 2021.

ANNEXURE C

PUBLICATIONS

Ayswaryalakshmi R, **"An Explainable Hybrid Deep Learning Framework for Facial Image Forgery Detection"**, attended the International Conference on SmartComputing, Artificial Intelligence and Data Engineering (SCAD 2026) held on 16th March 2026 at Sree Buddha College of Engineering, Pattoor, Alappuzha, Kerala.



SREE BUDDHA COLLEGE OF ENGINEERING (AUTONOMOUS)
PATTOOR, ALAPPUZHA, KERALA

CERTIFICATE
OF PARTICIPATION

This is to certify that

AYSWARYALAKSHMI R

has presented a paper titled

An Explainable Hybrid Deep Learning Framework for Facial Image Forgery Detection

in the International Conference on **S**mart **C**omputing, **A**rtificial Intelligence and **D**ata Engineering (**SCAD 2026**)
organized by Department of Computer Science and Engineering
and Department of Artificial Intelligence & Machine Learning
Sree Buddha College of Engineering Autonomous, Pattoor on 16th March , 2026.

SCAD 2026

Dr. Chinchu M S
Organizing Secretary, SCAD

Dr. S V Annlin Jeba
Convenor, SCAD

Dr. S Suresh Babu
Dean (Academics)

VISION & MISSION OF THE DEPARTMENT

VISION

- To evolve as a school of computing with globally reputed Centre's of excellence and serve the changing needs of the industry and society.

MISSION

- The department is committed in bringing out career-oriented graduates who are industry ready through innovative practices of teaching and learning process
- To nurture professional approach, leadership qualities and moral values to the graduates by organizing various programs periodically
- To acquire self-sustainability and serve the society through research and consultancy