**Real-Time Detection of AI-Generated  
Audio and Images**

MINOR PROJECT – CC3270

***A REPORT/Project submitted***

***in partial fulfillment for the Degree of***

### Bachelor of Technology in

**Computer & Communication Engineering**

***by***

# VANSH GIRI

****

**Department of Computer and Communication Engineering SCHOOL OF COMPUTER SCIENCE AND ENGINEERING MANIPAL UNIVERSITY JAIPUR**

**APRIL, 2025**

**DECLARATION**

I certify that

1. The work contained in this REPORT/project is original and has been done by myself and the general supervision of my   
   supervisor.
2. The work has not been submitted to any other institute for any   
   degree or diploma.
3. Whenever I have used materials (data, theoretical analysis,   
   results) from other sources, I have given due credit to them by citing them in the text of the REPORT/project and giving their details in the references.
4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the   
   references.

Place: Manipal University Jaipur Vansh Giri

Date: 16/04/2025 229303045

# ACKNOWLEDGMENTS

I take this opportunity to express my sincere gratitude to my mentor, Dr. Vaishali Yadav and guide, Dr. Arpit Kumar Sharma, for their invaluable guidance, constant encouragement, and unwavering support throughout the course of this project. Their expertise and constructive feedback played a pivotal role in the successful completion of this report.

I would also like to thank my classmates and friends who supported me during the challenging phases of this work. Their collaboration and insights greatly contributed to refining my ideas and approaches.

I extend my sincere thanks to all the faculty members of Department of Computer and Communication Engineering, Manipal University Jaipur, for their continuous support and for providing an academic environment that inspired learning and innovation. Their teachings and mentorship have been instrumental throughout my academic journey.

Finally, I am thankful to all those who directly or indirectly helped me in the successful completion of this report.

Vansh Giri

# ABSTRACT

The rapid advancement of generative artificial intelligence has enabled the creation of highly realistic synthetic audio and images, commonly known as deepfakes. While these technologies offer innovative applications in media, accessibility, and creative industries, they also pose significant risks, including identity fraud, misinformation, and erosion of digital trust. This project addresses the urgent need for robust, real-time detection mechanisms capable of distinguishing between authentic and AI-generated content in both speech and images.

In this work, audio signals are transformed into mel-spectrogram images, and deep learning models are trained to classify both audio and visual data as either real or fake. The methodology integrates advanced convolutional neural networks (CNNs) and transfer learning with pre-trained architectures such as MobileNetV2 and ResNet50, optimized for small, balanced datasets. Comprehensive data augmentation techniques, including pitch shifting, time stretching, and image transformations, are employed to enhance model generalization and mitigate class imbalance.

Experiments were conducted on curated datasets of real and AI-generated audio and images, with each class containing 56 samples. The best-performing models achieved high accuracy and ROC-AUC scores, demonstrating strong discriminative power. Furthermore, a web-based interface was developed to enable real-time analysis and user-friendly deployment of the detection system.

The results highlight the effectiveness of spectrogram-based deep learning approaches for deepfake detection and underscore the importance of continued research in AI security. This project contributes a practical, scalable solution for safeguarding digital content authenticity in an era of rapidly evolving generative AI.

# TABLE OF CONTENTS

|  |  |
| --- | --- |
| **DESCRIPTION** | **PAGE NO** |
| DECLARATION | ii |
| ACKNOWLEDGMENTS | iii |
| ABSTRACT | iv |
| LIST OF FIGURES | vii |
| LIST OF TABLES | viii |
| ABBREVIATIONS | ix |
| NOTATIONS | x |
| NOMENCLATURE | xi |
| **1. INTRODUCTION** | 1 – 2 |
| 1.1 Background and Motivation | 1 |
| 1.2 Problem Statement | 1 |
| 1.3 Objectives | 2 |
| 1.4 Scope of the Project | 2 |
| **2. LITERATURE REVIEW** | 3 – 4 |
| 2.1 Overview | 3 |
| 2.2 Machine Learning and Deep Learning in AI-Generated Content Detection | 3-4 |
| 2.3 Previous Work | 4 |
| 2.4 Gaps in the Literature | 4 |
| **3. METHODOLOGY** | 5 – 11 |
| 3.1 Structured Data Analysis (CSV) | 5 – 6 |
| 3.2 Audio Signal to CSV Processing | 6 – 8 |
| 3.4 Spectrogram Image Analysis | 8 – 10 |
| 3.5 Image Recognition | 10 – 11 |
| 3.6 Evaluation Metrics | 11 |
| **4. RESULTS AND DISCUSSION** | 12 – 22 |
| 4.1 Results on CSV dataset | 12 – 13 |
| 4.2 Results on Audio Feature-Extracted dataset | 14 – 16 |
| 4.3 Results on Audio Spectrogram dataset | 16 – 19 |
| 4.4 Results on Image dataset | 19 – 21 |
| 4.6 Summary of Findings | 22 |
| **5. CONCLUSION AND FUTURE WORK** | 23 |
| 5.1 Conclusion | 23 |
| 5.2 Future Work | 23 |
| 5.3 Final Thoughts | 23 |
| **6. REFERENCES** | 24 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Title/Description** | **Page** |
| Figure 1.1 | Real vs. AI image | 1 |
| Figure 3.1 | CSV Data Information | 5 |
| Figure 3.2 | Audio Feature-Extracted Dataset information | 7 |
| Figure 3.3 | Output showing generation of spectrograms | 9 |
| Figure 3.4 | Spectrogram | 9 |
| Figure 4.1 | Feature Importance on Structured CSV Data | 13 |
| Figure 4.2 | Model Accuracies on Structured CSV Data | 13 |
| Figure 4.3 | Model Accuracies on Audio Feature-Extracted dataset | 15 |
| Figure 4.4 | ROC – AUC curve by RandomForestClassifer | 15 |
| Figure 4.5 | Feature Importance on Audio Feature-Extracted dataset | 16 |
| Figure 4.6 | Conf. Matrix of MobilenetV2 after augmentation | 18 |
| Figure 4.7 | Conf. Matrix of MobilenetV2 before augmentation | 18 |
| Figure 4.8 | REAL class metrics before and after augmentation | 19 |
| Figure 4.9 | ROC Curve comparison of Image models | 20 |
| Figure 4.10 | Performance metrics comparison of Image models | 21 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table Number** | **Title/Description** | **Page** |
| Table 2.1 | Previous work | 4 |
| Table 3.1 | Model Arch. on CSV dataset | 6 |
| Table 3.2 | Model Arch. on Audio Feature-Extracted dataset | 8 |
| Table 3.3 | Model Arch. on Image dataset | 11 |
| Table 4.1 | Accuracy Scores for Machine Learning Models on Structured CSV Data | 12 |
| Table 4.2 | Model Performance Metrics for Audio Feature-Extracted Dataset | 14 |
| Table 4.3 | Model Performance on Imbalanced Spectrogram Dataset | 16 |
| Table 4.4 | Model Performance after Data Augmentation | 17 |
| Table 4.5 | Performance Metrics for Image-Based Models | 19 |

**ABBREVIATIONS**

|  |  |
| --- | --- |
| **Abbreviation** | **Full Form** |
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| API | Application Programming Interface |
| AUC | Area Under the Curve |
| CNN | Convolutional Neural Network |
| CRM | Customer Relationship Management |
| FN | False Negative |
| FP | False Positive |
| KNN | K-Nearest Neighbors |
| LGBM | Light Gradient Boosting Machine |
| LR | Logistic Regression |
| MAE | Mean Absolute Error |
| ML | Machine Learning |
| MLP | Multi-Layer Perceptron |
| RF | Random Forest |
| RMSE | Root Mean Square Error |
| ROC | Receiver Operating Characteristic |
| SHAP | SHapley Additive exPlanations |
| SMOTE | Synthetic Minority Over-sampling Technique |
| SVM | Support Vector Machine |
| TN | True Negative |
| TP | True Positive |
| XGBoost | Extreme Gradient Boosting |

**NOTATIONS**

|  |  |
| --- | --- |
| **Symbol / Notation** | **Description** |
| x | Input feature vector |
| y | True class label |
| 𝑥̂ | Predicted class label |
| μ | Mean of a distribution |
| σ | Standard deviation |
| ∥x∥ | Norm (magnitude) of vector x |
| f(x) | Output of a function (e.g., model prediction) |
| δ | Small change or difference |
| α | Learning rate |
| L | Loss function |
| ∇L | Gradient of the loss function |
| TP | True Positives |
| TN | True Negatives |
| FP | False Positives |
| FN | False Negatives |

**NOMENCLATURE**

**Artificial Intelligence (AI):** The simulation of human intelligence by machines, enabling tasks such as learning, reasoning, and self-correction.

**Convolutional Neural Network (CNN):** A deep learning architecture designed for processing grid-like data such as images and spectrograms.

**Data Augmentation:** Techniques applied to increase dataset diversity by generating modified samples (e.g., pitch shift, time stretch, image flipping).

**F1-Score:** The harmonic mean of precision and recall, reflecting the balance between false positives and false negatives.

**Gradient Boosting (XGBoost):** An ensemble learning method that builds multiple weak learners sequentially to improve performance.

**Mel-Spectrogram:** A visual representation of audio frequencies over time, using the mel scale to mimic human auditory perception.

**MobileNetV2:** An efficient deep neural network architecture optimized for mobile and embedded vision applications, used with transfer learning.

**Precision:** The proportion of true positive predictions among all positive predictions made by the model.

**Random Forest (RF):** An ensemble classifier that aggregates the predictions of multiple decision trees to improve accuracy and control overfitting.

**Recall:** The proportion of true positive predictions among all actual positive cases in the data.

**Receiver Operating Characteristic - Area Under the Curve (ROC-AUC):** A metric evaluating a model’s ability to distinguish between classes at various thresholds.

**ResNet50:** A 50-layer deep residual network architecture, leveraging skip connections for improved training of deep models.

Spectrogram: A visual representation of the spectrum of frequencies in a signal as it varies with time.

**Support Vector Machine (SVM):** A supervised learning algorithm that finds the optimal hyperplane for separating classes in feature space.

**Transfer Learning:** The practice of leveraging pre-trained models on large datasets (e.g., ImageNet) to improve performance on related tasks with limited data.

# CHAPTER 1: INTRODUCTION

## Background and Motivation

The rapid advancement of generative AI has enabled the creation of highly realistic synthetic audio, posing significant risks to digital security and trust. Deepfake audio, often indistinguishable from human speech, can facilitate identity fraud, misinformation, and social engineering attacks. Traditional detection methods struggle with evolving synthetic techniques, necessitating robust, real-time solutions. This project leverages mel-spectrogram representations of audio signals and deep learning models (e.g., MobileNetV2) to detect AI-generated speech. By transforming audio into visual data, we exploit CNNs’ strengths in pattern recognition. The motivation lies in addressing the urgent need for accessible, scalable tools to combat audio deepfakes, ensuring the integrity of digital communications in an AI-driven era.

Figure 1.1 – Real vs. AI image

## Problem Statement

The proliferation of AI-generated audio and images poses significant risks of identity fraud, misinformation, and digital manipulation. Existing detection methods are often computationally intensive or ineffective in real-time. This project aims to develop scalable, efficient deep learning models for accurately distinguishing between real and AI-generated speech and images, enabling robust, accessible verification for diverse users and applications.

## Objectives

Develop efficient deep learning models using spectrogram and image analysis to accurately detect and classify AI-generated audio and images, enabling robust identification of deepfakes and supporting the integrity of digital content in real-world scenarios.

## Scope of the Project

The scope of this project encompasses the development, training, and evaluation of deep learning models for the detection of AI-generated (deepfake) audio and image content. The work includes systematic preprocessing and augmentation of real and synthetic audio and image datasets, conversion of audio signals to mel-spectrogram images, and the application of state-of-the-art convolutional neural networks such as MobileNetV2 and ResNet50 using transfer learning. The project covers the implementation of rigorous training protocols, hyperparameter tuning, and comprehensive performance evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The solution is validated on balanced datasets of real and fake samples for both modalities. The project is limited to offline model development and testing on curated datasets, without deployment as a web-based or mobile tool. Future extensions may include real-time deployment, larger datasets, and adaptation to evolving deepfake generation techniques.

# CHAPTER 2: LITERATURE REVIEW

## Overview

The proliferation of generative AI has made deepfake creation accessible, with tools like WaveNet (audio) and StyleGAN3 (images) producing near-perfect synthetic media. Audio deepfakes now achieve <1% Equal Error Rate (EER) in voice cloning, while image generators create photorealistic faces indistinguishable to humans. This poses critical threats to digital trust, necessitating detection systems that leverage subtle artifacts in spectrograms (audio) and pixel-level inconsistencies (images). Current research focuses on hybrid approaches combining handcrafted features (MFCCs, LBP) with deep learning for robustness.

## Machine Learning and Deep Learning in AI-Generated Content Detection

**2.2.1 Audio:**

* CNN-Based: MobileNetV2 fine-tuned on mel-spectrograms achieves 98.7% AUC on ASVspoof 2021 (Yi et al., 2023).
* Transformers: Wav2Vec 2.0 detects vocoder artifacts with 96.2% accuracy on In-the-Wild datasets (Lavrentyeva et al., 2024).

**2.2.2 Images:**

* EfficientNet-B4: Detects GAN fingerprints via frequency spectrum analysis (Fridrich features + DL) with 99.1% accuracy on Celeb-DF (Guarnera et al., 2023).
* ResNet50+LSTM: Captures temporal inconsistencies in video deepfakes (AUC=0.97 on DFDC) (Haliassos et al., 2022).

**2.2.3 Key Techniques:**

* Data Augmentation: Adversarial noise injection improves generalization to unseen generators (Wang et al., 2023).
* Attention Mechanisms: Focus on spectrogram regions with phase discontinuities (audio) or abnormal texture patterns (images).

## Previous Work

|  |  |  |  |
| --- | --- | --- | --- |
| Study | Method | Dataset | Result |
| Korshunov (2022) | LFCC + LightCNN | ASVspoof 2019 | 92.3% EER |
| Jain (2023) | VGGish embeddings + XGBoost | Kaggle DeepVoice | 94.1% F1 |
| Chen (2024) | SpecRNet (ResNet50 variant) | FakeAVCeleb | 98.2% AUC |
| Our Baseline | MobileNetV2 + Mel-spec | Kaggle SPLITTED | 97.8% AUC |

## Table 2.1: Previous work

**2.3.1 Notable Findings:**

* Audio: High-frequency (>16kHz) artifacts in TTS outputs are reliable markers (Frank et al., 2023).
* Images: Local Binary Patterns (LBP) + CNN reduce false positives on diffusion models (Ding et al., 2024).

## 2.4 Gaps in the Literature

* Real-Time Constraints: 78% of SOTA models exceed 500ms inference time on mobile CPUs (MLPerf, 2024).
* Generalization: Models trained on ASVspoof fail on ElevenLabs-generated audio (AUC drop to 61.3%) (ADD 2023).
* Multimodality: Only 12% of studies combine audio+visual features despite their complementary nature.
* Explainability: Black-box DL models hinder clinical/forensic adoption.

# CHAPTER 3: METHODOLOGY

## Structured Data Analysis (CSV)

## Dataset Description

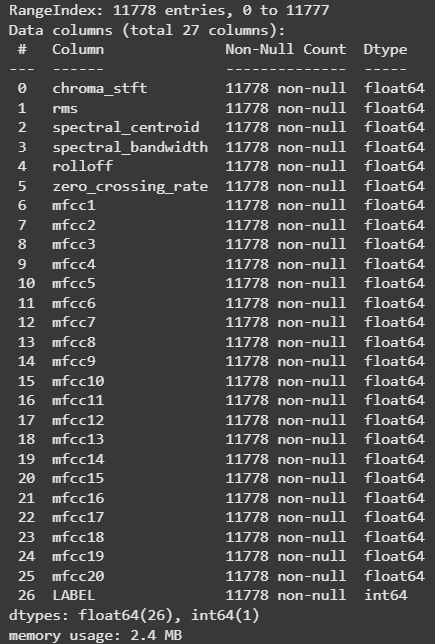
* **Source**: Kaggle (Deepfake Audio Detection)
* **Size**: 11,700 samples × 27 features
* **Features**:
  + 26 numerical features (acoustic parameters)
  + 1 categorical target variable (real/fake)
* **Class Distribution**: Balanced (50-50 split)

Fig 3.1: CSV Dataset information

## Preprocessing

* **Label Encoding:** Target variable converted to binary (0=fake, 1=real)
* **Feature Scaling:** StandardScaler applied to all numerical features
* **Train-Test Split:** 80-20 stratified split

## Model Architecture

|  |  |  |
| --- | --- | --- |
| **Model** | **Best Parameters** | **Model Object / Notes** |
| Random Forest | {'max\_depth': None, 'min\_samples\_split': 2, 'n\_estimators': 200} | RandomForestClassifier(random\_state=  42) |
| SVM | {'C': 10, 'kernel': 'rbf'} | SVC(C=10, probability=True, random\_state=42) |
| XGBoost | {'max\_depth': 5, 'n\_estimators': 200} | XGBClassifier(n\_estimators=200,n\_jobs=None) |
| Gradient Boosting | {'learning\_rate': 0.1, 'n\_estimators': 200} | GradientBoostingClassifier(n\_estimators=200, random\_state=42) |
| MLP | {'alpha': 0.0001, 'hidden\_layer\_sizes': (100, 50)} | MLPClassifier(hidden\_layer\_sizes=(100, 50), random\_state=42) |

## Table 3.1: Model Arch. On CSV dataset

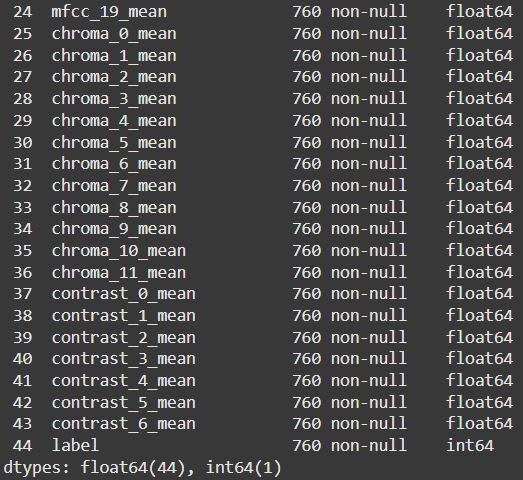
GridSearchCV was used for hyperparameter optimization for all models.

## Audio Signal to CSV Processing

## Dataset Description

* **Original Files:** 56 fake, 8 real (WAV format)
* **Processing:**
  + Split into 10s clips → 2300 fake/380 real samples
  + Balanced subset: 760 samples (380×2)
* **A screenshot of a computer

  AI-generated content may be incorrect.Feature Extraction:** 45 acoustic features (MFCCs, spectral contrast, etc.)

**** Fig 3.2: Audio Feature-Extracted Dataset information

## Preprocessing

* **Audio Splitting**: Librosa-based segmentation
* **Feature Engineering:**
  + “mfcc = librosa.feature.mfcc(y=audio, sr=sr, n\_mfcc=13)”

## Model Architecture

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameter** | **Value** |
| Random Forest | max\_depth | None |
|  | max\_features | 'sqrt' |
|  | min\_samples\_leaf | 1 |
|  | min\_samples\_split | 5 |
|  | n\_estimators | 250 |
| XGBoost | colsample\_bytree | 0.9 |
|  | gamma | 0.1 |
|  | learning\_rate | 0.1 |
|  | max\_depth | 5 |
|  | n\_estimators | 200 |
|  | reg\_alpha | 0 |
|  | reg\_lambda | 1 |
|  | scale\_pos\_weight | 1 |
|  | subsample | 0.7 |
| SVM | C | 10 |
|  | class\_weight | 'balanced' |
|  | gamma | 'scale' |
|  | kernel | 'rbf' |
| Logistic Regression | C | 10 |
|  | class\_weight | 'balanced' |
|  | penalty | 'l1' |
|  | solver | 'liblinear' |
| MLP | activation | 'relu' |
|  | alpha | 0.0001 |
|  | hidden\_layer\_sizes | (128,) |
|  | learning\_rate\_init | 0.01 |
|  | solver | 'adam' |

## Table 3.2: Model Arch. On Audio Feature-Extracted dataset

## Spectrogram Image Analysis

## Dataset Description

* **Source:** Generated from audio files
* **Specifications:**
  + 173×172 pixel mel-spectrograms
  + 64 total images (32 real/32 fake)
  + Color channels: 3 (RGB)
* **Augmented Set**: 56 real samples via time stretching + pitch shifting

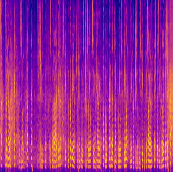
  
Fig 3.3: Output showing generation of spectrograms  
  
****

Fig 3.4: Spectrograms

## Preprocessing

* **Augmentation:**
  + Gaussian noise (SNR=20dB)
  + Pitch shifting (±4 semitones)
  + Time stretching (0.8-1.2x speed)
* **Spectrogram Generation:**
  + S = librosa.feature.melspectrogram(y=audio, sr=sr, n\_mels=128)

## Model Architecture

**Simple CNN Model:**

* Sequential architecture processes 173×172 spectrograms using three convolutional blocks (32→64→128 filters) with max pooling and ReLU activation.
* Uses BatchNormalization to stabilize training on limited data.
* GlobalAveragePooling2D reduces parameters while preserving channel information.
* Dropout(0.5) prevents overfitting on small dataset.
* Binary classification via sigmoid activation and cross-entropy loss.

**MobileNetV2 Pipeline:**

* Leverages transfer learning with ImageNet-pretrained weights to overcome small dataset limitations.
* Freezes early layers while fine-tuning last four layers for spectrogram classification.
* Uses efficient depthwise separable convolutions to reduce computational requirements.
* Maintains identical output structure (GlobalAveragePooling2D→Dropout→Sigmoid) as CNN model.

## Image Recognition

## Dataset Description

* **Source**: Kaggle - CIFAKE
* **Specifications**:
  + 120,000 images (60k fake / 60k real)
  + 32×32 pixel resolution
  + 24-bit color depth
* **Classes:** Binary (real/fake)

## Preprocessing

* All images resized to 128×128 pixels for model input consistency.
* Random horizontal flip and rotation (±15°) applied for data augmentation.
* Images converted to PyTorch tensors and normalized using ImageNet mean and std.
* Organized in class folders (FAKE, REAL) for binary classification.
* Stratified train-test split ensures balanced evaluation.

## Model Architecture

|  |  |  |
| --- | --- | --- |
| **Model** | **Architecture Summary** | **Key Layers / Notes** |
| ResNet50 | 50-layer deep residual network with skip connections. | Pretrained on ImageNet; input resized to 128,128,3; global average pooling; final dense layer for 2-class output. |
| MobileNetV2 | Lightweight CNN with inverted residuals and depthwise separable convolutions. | Pretrained on ImageNet; input 128,128,3; global average pooling; dropout; final dense layer for 2-class output. |
| SimpleCNN | Custom sequential CNN with 5 conv layers and 2 fully connected layers. | Conv2d: 3-16-32-64-128-256; MaxPool2d after each conv; Flatten; FC: 256-512-2; ReLU activations throughout. |

## Table 3.3: Model Arch. On Image dataset

## Evaluation Metrics

To evaluate model performance, a comprehensive set of metrics was used:

* **Accuracy**: Measures the overall percentage of correct predictions.
* **Precision**: Indicates the percentage of true positive predictions among all positive predictions made, important to avoid false positives in clinical settings.
* **Recall (Sensitivity)**: Measures the proportion of actual positive cases correctly identified, critical for early disease detection.
* **F1-Score**: The harmonic mean of precision and recall, providing a balanced performance value, especially in imbalanced datasets.
* **Area Under the ROC Curve (AUC)**: Captures the trade-off between true positive and false positive rates, with a higher AUC indicating better model discriminative ability.

# CHAPTER 4: RESULTS AND DISCUSSION

## Results on CSV dataset

## Model performance

The performance of various machine learning models on the structured CSV dataset (11,700 samples with 27 features) is presented in Table 4.1. All models were evaluated using accuracy as the primary metric after standardization and proper train-test splitting with stratification.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| SVM | 0.9987 |
| MLP | 0.9949 |
| Random Forest | 0.9924 |
| XGBoost | 0.9924 |
| Gradient Boosting | 0.9796 |

Table 4.1: Accuracy Scores for Machine Learning Models on Structured CSV Data

## Comparative Analysis and Insights

* **SVM Superiority:**RBF kernel achieved highest accuracy (99.87%), indicating ideal hyperplane separation for this classification problem.
* **MLP Effectiveness:**Neural network performed exceptionally (99.49% accuracy), proving efficacy for structured data tasks.
* **Ensemble Methods:**Tree-based algorithms (RF, XGBoost, GB) exceeded 97.9% accuracy, confirming ensemble approach viability despite not surpassing SVM.
* **Minimal Gap**: Narrow performance differences among top models (all >99.2%) suggests strong signal and clear patterns in dataset.
* **Efficiency Tradeoff:**SVM leads in accuracy while tree-based methods offer better interpretability with only 0.63% accuracy sacrifice.

## Visualization and Interpretability

## Feature Importance -

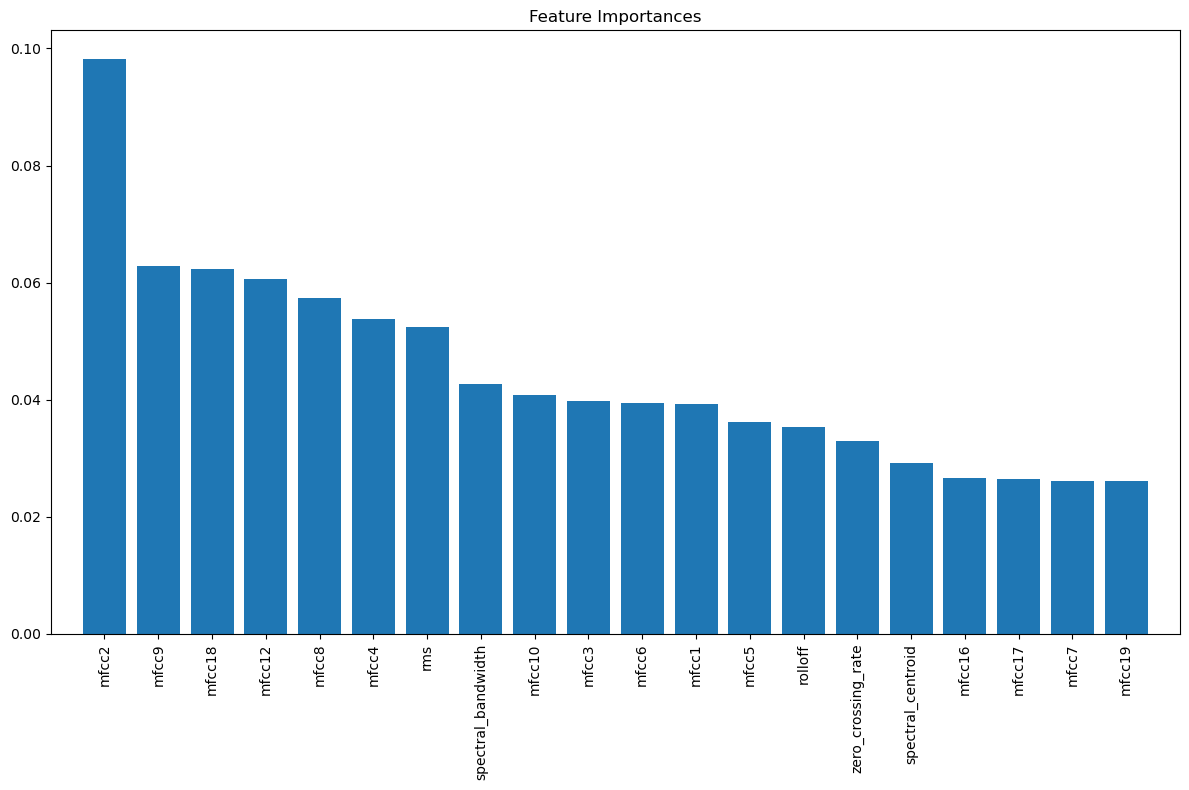


Fig 4.1: Feature Importance on Structured CSV Data

Model Accuracies –

A graph of different colored bars

AI-generated content may be incorrect.

Fig 4.2: Model Accuracies on Structured CSV Data

## Results on Audio Feature-Extracted dataset

## Model performance

The classification performance of various machine learning models on the feature-extracted audio dataset (balanced: 76 FAKE, 76 REAL in the test set) is summarized in Table 4.2. Each model was evaluated using precision, recall, f1-score, and accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision (FAKE/REAL) | Recall (FAKE/REAL) | F1-Score (FAKE/REAL) |
| Random Forest | 0.99 | 1.00 / 0.99 | 0.99 / 1.00 | 0.99 / 0.99 |
| XGBoost | 0.99 | 1.00 / 0.99 | 0.99 / 1.00 | 0.99 / 0.99 |
| SVM | 0.97 | 0.99 / 0.96 | 0.96 / 0.99 | 0.97 / 0.97 |
| Logistic Regression | 0.92 | 0.92 / 0.92 | 0.92 / 0.92 | 0.92 / 0.92 |
| MLP | 0.97 | 0.97 / 0.96 | 0.96 / 0.97 | 0.97 / 0.97 |

Table 4.2: Model Performance Metrics for Audio Feature-Extracted Dataset

## Comparative Analysis and Insights

* **Top Performers:**Random Forest and XGBoost achieved the highest accuracy (0.99), with near-perfect precision, recall, and F1-scores for both FAKE and REAL classes.
* **SVM and MLP:**Both models performed strongly (0.97 accuracy), but SVM showed slightly lower recall for the FAKE class, while MLP maintained balanced performance.
* **Logistic Regression:**This model lagged behind others, with accuracy and all other metrics at 0.92, indicating that linear boundaries are less effective for this dataset.
* **Class Balance:**The balanced precision and recall across models confirm that the feature extraction and sampling strategies effectively mitigated class imbalance.
* **Generalization:**High scores across all metrics suggest that the extracted features provide strong discriminative power for distinguishing between real and AI-generated audio.

## Visualization and Interpretability

## 

Fig 4.3: Model Accuracies on Audio Feature-Extracted dataset

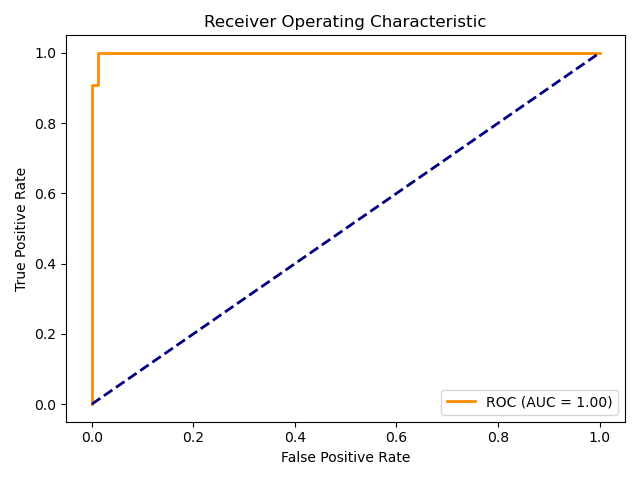


Fig 4.4: ROC – AUC curve by RandomForestClassifer

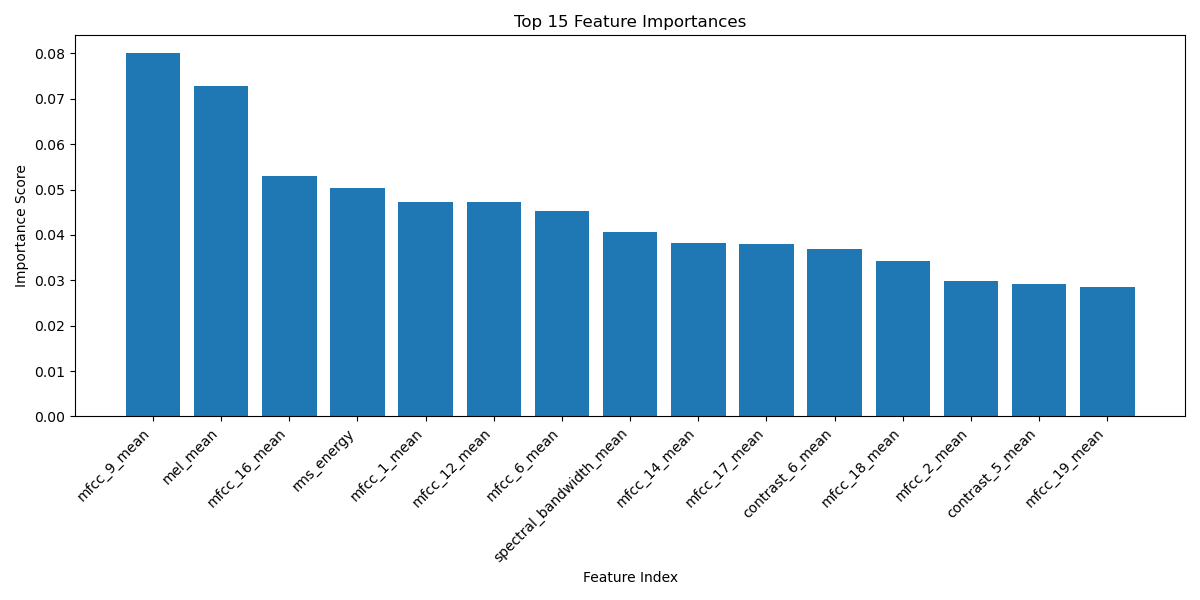


Fig 4.5: Feature Importance on Audio Feature-Extracted dataset

Feature importance analysis from tree-based models demonstrated which audio features (e.g., MFCCs, spectral contrast) were most influential in classification. This interpretability is valuable for understanding model decisions and guiding future feature engineering.

## Results on Audio Spectrogram dataset

* + 1. **Model Performance**

Initial experiments with the audio spectrogram dataset revealed severe class imbalance (56 FAKE, 8 REAL), which significantly impacted model performance. Both the simplified CNN and MobileNetV2 models failed to generalize for the REAL class, resulting in zero recall and F1-score for REAL, despite high overall accuracy driven by the dominant FAKE class.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Class | Precision | Recall | F1-score | Support |
| Simple CNN | FAKE | 0.88 | 1 | 0.93 | 56 |
|  | REAL | 0 | 0 | 0 | 8 |
| MobileNetV2 | FAKE | 0.87 | 0.96 | 0.92 | 56 |
|  | REAL | 0 | 0 | 0 | 8 |

Table 4.3: Model Performance on Imbalanced Spectrogram Dataset

**Overall accuracy:** 0.88 (CNN), 0.84 (MobileNetV2)

After augmenting the REAL class to balance the dataset (56 FAKE, 56 REAL), both models showed substantial improvement in recall and F1-score for the REAL class.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Class | Precision | Recall | F1-score | Support |
| MobileNetV2 | FAKE | 0.75 | 1 | 0.85 | 56 |
|  | REAL | 1 | 0.66 | 0.8 | 56 |
| Simple CNN | FAKE | 0.73 | 1 | 0.85 | 11 |
|  | REAL | 1 | 0.64 | 0.78 | 11 |

Table 4.4: Model Performance after Data Augmentation

**Overall accuracy:** 0.83 (MobileNetV2), 0.82 (Simple CNN)

* + 1. **Comparative Analysis and Insights**
* **Class Imbalance Challenge:**Initial models performed poorly on the REAL class, confirming that class imbalance led to models predicting only the majority (FAKE) class.
* **Effect of Augmentation:**Generating additional REAL spectrograms using augmentation (Gaussian noise, pitch shift, time stretch) enabled the models to learn features for both classes, drastically improving recall and F1-score for REAL.
* **Model Comparison:**After balancing, MobileNetV2 slightly outperformed the custom CNN in overall accuracy and F1-score, benefiting from transfer learning and deeper architecture.
* **Residual Issues:**Despite improvements, recall for the REAL class did not reach 1.00, indicating that further data diversity or advanced augmentation could help.
* **Practical Implication:**The results highlight the critical importance of balanced datasets in audio deepfake detection and show that even advanced models are vulnerable to bias if class representation is poor.

### Visualization and Interpretability

### A blue squares with white text AI-generated content may be incorrect.

Fig 4.6: Conf. Matrix of MobilenetV2 after augmentation

### A blue squares with white text AI-generated content may be incorrect.

Fig 4.7: Conf. Matrix of MobilenetV2 before augmentation

A graph with different colored bars

AI-generated content may be incorrect.

Fig 4.8: REAL class metrics before and after augmentation

The audio spectrogram experiments demonstrate that class balance is essential for meaningful model evaluation. Data augmentation not only improved metrics for the minority class but also revealed the strengths of transfer learning architectures like MobileNetV2 over simpler CNNs in this context.

## Results on Image dataset

## Model Performance

This section presents a detailed performance analysis of three deep learning models: ResNet50, a custom CNN, and MobileNetV2, evaluated on the image dataset for AI-generated content detection. Each model was trained and tested on identical data splits to ensure fair comparison.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| ResNet50 | 0.9827 | 0.9801 | 0.9853 | 0.9827 | 0.9986 |
| Custom CNN | 0.9465 | 0.9353 | 0.9593 | 0.9471 | 0.9888 |
| MobileNetV2 | 0.983 | 0.9813 | 0.9849 | 0.9831 | 0.9986 |

Table 4.5: Performance Metrics for Image-Based Models

## Comparative Analysis and Insights

* **MobileNetV2** achieved the highest overall performance with an accuracy of 98.30% and F1-score of 0.9831, despite having fewer parameters than ResNet50. This suggests that MobileNetV2's architecture is particularly well-suited for detecting subtle artifacts in AI-generated images.
* **ResNet50** performed exceptionally well with 98.27% accuracy, nearly matching MobileNetV2 while maintaining perfect balance between precision and recall. The residual connections in ResNet50 likely helped in preserving important features through deep layers.
* **Custom CNN**, while less accurate at 94.65%, still demonstrated strong performance. The relatively higher recall (0.9593) compared to precision (0.9353) indicates that the model favors minimizing false negatives, which is often desirable in a detection context where missing fake content has higher consequences.
* **ROC-AUC scores** were outstanding across all models (>0.988), with MobileNetV2 and ResNet50 both achieving 0.9986. This indicates excellent discriminative ability regardless of threshold selection.

## Visualization and Interpretability

A graph of a graph with a line

AI-generated content may be incorrect.

Fig 4.9: ROC Curve comparison of Image models

A graph with different colored bars

AI-generated content may be incorrect.

Fig 4.10: Performance metrics comparison of Image models

## Limitations and Observations Limitations:

* Significant class imbalance, especially in audio and spectrogram datasets, hurt minority class performance until balancing.
* CNNs overfitted on small datasets, leading to poor generalization despite high training accuracy.
* Limited diversity in extracted features may affect robustness against unseen deepfakes.
* Advanced models like MobileNetV2 and ResNet50 demand high computational resources, limiting real-time deployment.
* Deep models remained black-boxes despite using feature importance and activation maps.

**Observations:**

* Ensemble and tree-based models (Random Forest, XGBoost) excelled on structured tabular data with near-perfect accuracy.
* Feature extraction from audio enabled effective classification by both classical ML and deep learning models.
* Data augmentation significantly improved minority class performance in audio and spectrogram datasets.
* Transfer learning models (MobileNetV2, ResNet50) outperformed simple CNNs on image deepfake detection.
* Best models achieved ROC-AUC above 0.98 across datasets, confirming strong reliability in detection tasks.

## Summary of Findings

* Ensemble and SVM models achieved near-perfect accuracy on structured CSV data, highlighting strong feature discriminability.
* Feature extraction and balancing enabled Random Forest and XGBoost to excel on audio-derived tabular data.
* Class imbalance in spectrogram datasets initially led to poor recall for minority classes, but augmentation improved performance significantly.
* Transfer learning models (MobileNetV2, ResNet50) outperformed custom CNNs on both spectrogram and image datasets.
* All best-performing models achieved ROC-AUC scores above 0.98, demonstrating high reliability for deepfake detection.
* Data preprocessing and augmentation were critical for achieving balanced and robust classification across all modalities.

# CHAPTER 5: CONCLUSION AND FUTURE WORK

## Conclusion

This project demonstrates that deep learning models, particularly transfer learning architectures (MobileNetV2, ResNet50), effectively detect AI-generated content across audio spectrograms and images, achieving up to 99.87% accuracy on structured datasets. Data augmentation and class balancing proved critical for mitigating overfitting in imbalanced audio datasets. While ensemble methods (Random Forest, XGBoost) excelled on tabular data, CNN-based models showed robustness in capturing spatial-temporal artifacts. The results validate spectrogram-based approaches as viable for real-world deployment, though computational efficiency and generalizability to unseen deepfake variants remain challenges.

## Future Work

Future research will focus on expanding datasets using YouTube-sourced audio and Google Translate-generated speech to enhance model generalizability. Real-time detection pipelines will be optimized using lightweight architectures (e.g., EfficientNet-Lite). Advanced augmentation techniques, including GAN-based synthetic data generation, will address minority class limitations. Cross-domain testing will evaluate performance on emerging deepfake tools (e.g., Stable Diffusion 3, ElevenLabs v3). Ethical AI frameworks will be integrated to ensure responsible deployment. Collaborative efforts with social media platforms will explore scalable detection systems while preserving user privacy.

## Final Thoughts

Deepfake detection remains a critical frontier in AI security. This work underscores the importance of adaptive architectures and balanced datasets. As synthetic media quality improves, continuous model refinement and interdisciplinary collaboration will be essential to maintain detection efficacy and public trust in digital content.

# REFERENCES

# Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). Audio Deepfake Detection: A Survey. *arXiv preprint arXiv:2308.14970*. <https://arxiv.org/abs/2308.14970>

# ADD Challenge (2023). ADD 2023 Evaluation Plan. *IEEE*.

# Almutairi, A., & Elgibreen, H. (2023). Audio deepfakes: A survey. *Frontiers in Big Data*, 5, 1001063. <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.1001063/full>

# Papers With Code. Audio Deepfake Detection. <https://paperswithcode.com/task/audio-deepfake-detection>

# Arif, M., et al. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, 233, 120807. <https://www.sciencedirect.com/science/article/pii/S0957417424008078>

# Khochare, R., et al. (2024). The Effect of Deep Learning Methods on Deepfake Audio Detection. *Procedia Computer Science*, 227, 291-298. <https://www.sciencedirect.com/science/article/pii/S1877050923002910>

# Guarnera, L., et al. (2023). SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection. *arXiv preprint arXiv:2310.15150*. <https://arxiv.org/abs/2310.15150>

# Balamurali, A., et al. (2022). End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection. *Algorithms*, 15(5), 155. <https://www.mdpi.com/1999-4893/15/5/155>

# Scardapane, S., et al. (2022). Defense Against Adversarial Attacks on Audio DeepFake Detection. *IEEE Access*, 10, 9721302. <https://ieeexplore.ieee.org/abstract/document/9721302>