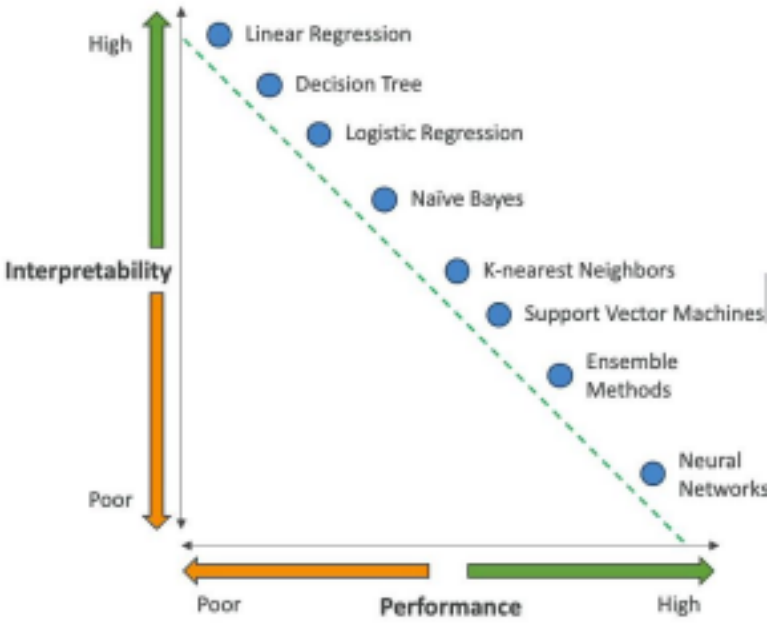
# Progress Report: Deepfake Audio Detection

*Hisham Iqbal Khokhar*

# 1. Introduction

In recent years, AI-generated voices have become increasingly sophisticated, particularly in applications such as virtual assistants, media production, and customer service automation. These AI voices are often generated using deep learning models, making it difficult to distinguish them from real human speech. This presents challenges in fields such as security, fraud detection, and content moderation. Motivated by these challenges, this project aims to develop a machine learning model capable of detecting AI-generated voices and differentiating them from human speech.

1.1. Objectives/Next steps: -




Based on the figure above depicting the performance vs inerpretability of common ML models, we can notice that models like linear regression and decision trees are highly interpretable but offer moderate performance, making them ideal for applications requiring transparency. As we move towards more complex models like ensemble methods and neural networks, performance improves significantly, but interpretability declines due to their complexity and ”black-box” nature. Considering this and based on the progress of the research , our objectives/next steps will be: -

1.1.1. Model Training:

- Use the 6 principal components as input features for machine learning models to classify audio into the four labels.

- Experiment with models such as Random Forest, Support Vector Machine (SVM), and Neural Networks.

1.1.2. Evaluation:
- Assess model performance using metrics like accuracy, precision, recall, and F1-score.
- Validate results using k-fold cross-validation to ensure robustness.

1.1.3. Further Improvements:
- Investigate feature engineering to extract additional unique characteristics.
- Explore oversampling techniques if class imbalance exists.

### 1.2. Product Overview

#### 1.2.1. Scope

The model will be developed to classify audio clips as either AI-generated or real human voices. It will process pre-labeled audio clips and use audio features such as pitch variability, spectral patterns, and prosody to make predictions. A wide range of AI generated audio of men and women with different accents, tones, pitch, speed have been taken into account to distinguish them between actual real audio.

#### 1.2.2. Audience

This project is intended for individuals and organizations concerned with the detection of synthetic voices, including professionals in content moderation, security, and fraud prevention. It may also be of interest to researchers in the field of voice recognition and artificial intelligence.

## 2. Background Review

### 2.1. Existing Approaches

Existing solutions for detecting AI-generated voices include machine learning models such as Support Vector Machines (SVMs), Random Forests, and neural networks applied to various audio features. While these traditional methods have shown some success, they often struggle to generalize across diverse platforms and evolving AI voice generation techniques. Recent research from 2019 to 2024 has focused on enhancing generalization through advanced deep learning models. However, existing methods still face challenges in distinguishing subtle differences between human speech and increasingly sophisticated

synthetic voices. This project addresses these challenges by leveraging deep learning techniques that integrate both temporal and spatial audio patterns for improved detection accuracy.

### 2.2. Related Literature

Several studies have explored voice synthesis and detection, with foundational work utilizing Mel Frequency Cepstral Coefficients (MFCCs), a widely used feature in speech analysis. Other research highlights the effectiveness of deep learning models, particularly Convolutional Neural Networks (CNNs), for classifying and distinguishing audio signals. However, CNNs and Recurrent Neural Networks (RNNs) often struggle to perform well when faced with unseen synthetic voices. More recent approaches have introduced transformer-based architectures, which offer improved context-awareness for processing audio data. Building on this prior work, this project proposes a hybrid model that combines CNNs and transformers to better capture spatial and temporal audio patterns. This approach aims to overcome the limitations of existing methods and enhance detection performance.

## 3. Dataset Overview

### 3.1. Initial Dataset:

2158 data points categorized into four labels: -
M_REAL: Real male audio, M_AI: AI-generated male audio, F_REAL: Real female audio F_AI: AI-generated female audio

### 3.2. Initial Features:

File Size (KB), Spectral Centroid Mean, Spectral Bandwidth Mean, RMS Mean, Zero-Crossing Rate Mean, Spectral Contrast Mean, Pitch Mean, Pitch Confidence Mean, Mel Spectrogram Mean, and Mel Spectrogram Variance.

### 3.3. Feature Augmentation:

Prosodic and temporal features were added:

- Energy Mean and Speech Rate.
- Missing data was imputed with the column mean to avoid data loss.

## 4. Preprocessing

### 4.1. Normalization:

All audio files were volume-normalized using Audacity to maintain uniform decibel levels across samples. Feature values were also scaled for consistency, enabling effective

comparisons and reducing bias during training.

## 5. Analysis of Feature Distributions

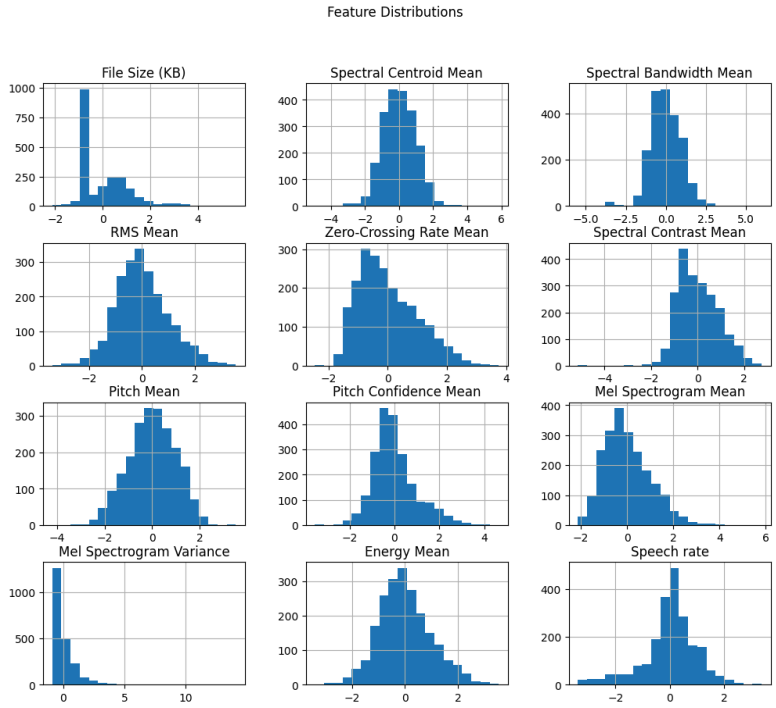
The attached histograms visualize the distributions of the extracted features. Key observations: -

1. File Size (KB):
The distribution is skewed heavily, indicating variability in audio duration or file encoding across samples.

2. Energy Mean, Speech Rate, RMS Mean:
These features exhibit normal distributions, centered around a mean with slight variance, suggesting consistent dynamic ranges in the dataset.

3. Mel Spectrogram Variance:
Shows a long-tail distribution, possibly highlighting outliers or specific file properties affecting variability in the spectrogram.

4. Other Features (Spectral, Pitch):
Spectral and pitch-related features exhibit Gaussian-like distributions, signifying good normalization and data integrity.



## 6. Correlation Matrix Analysis

### 6.1. High Positive Correlations:

Mel Spectrogram Mean and RMS Mean (0.93): Suggests that root mean square energy closely mirrors the spectrogram's average intensity.

Energy Mean and RMS Mean (0.93): Reinforces the link between average signal intensity and its computed energy.

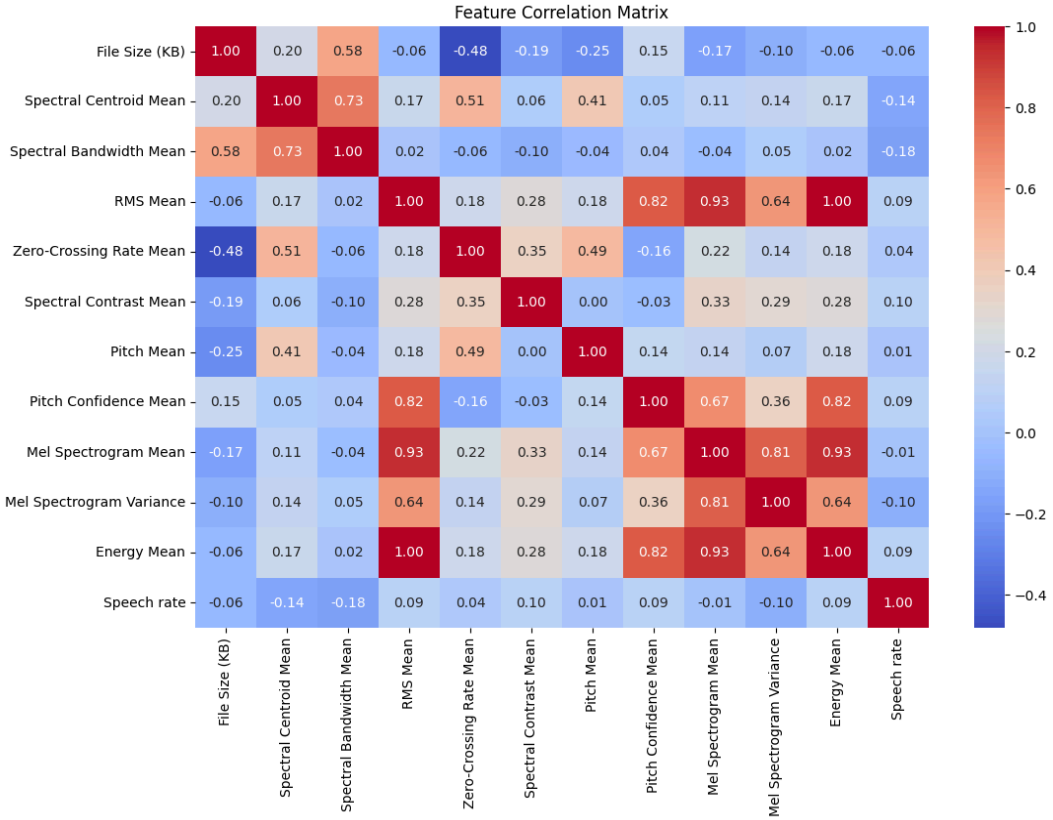
Pitch Confidence Mean and RMS Mean (0.82): Indicates that higher pitch stability correlates with stronger energy levels in audio.

### 6.2. Negative Correlations:

Zero-Crossing Rate Mean and File Size (-0.48): Smaller file sizes may correspond to simpler waveforms with more zero-crossings.

### 6.3. Weak Relationships:

Speech Rate exhibits minimal correlation with most features, potentially indicating independence and unique predictive value.



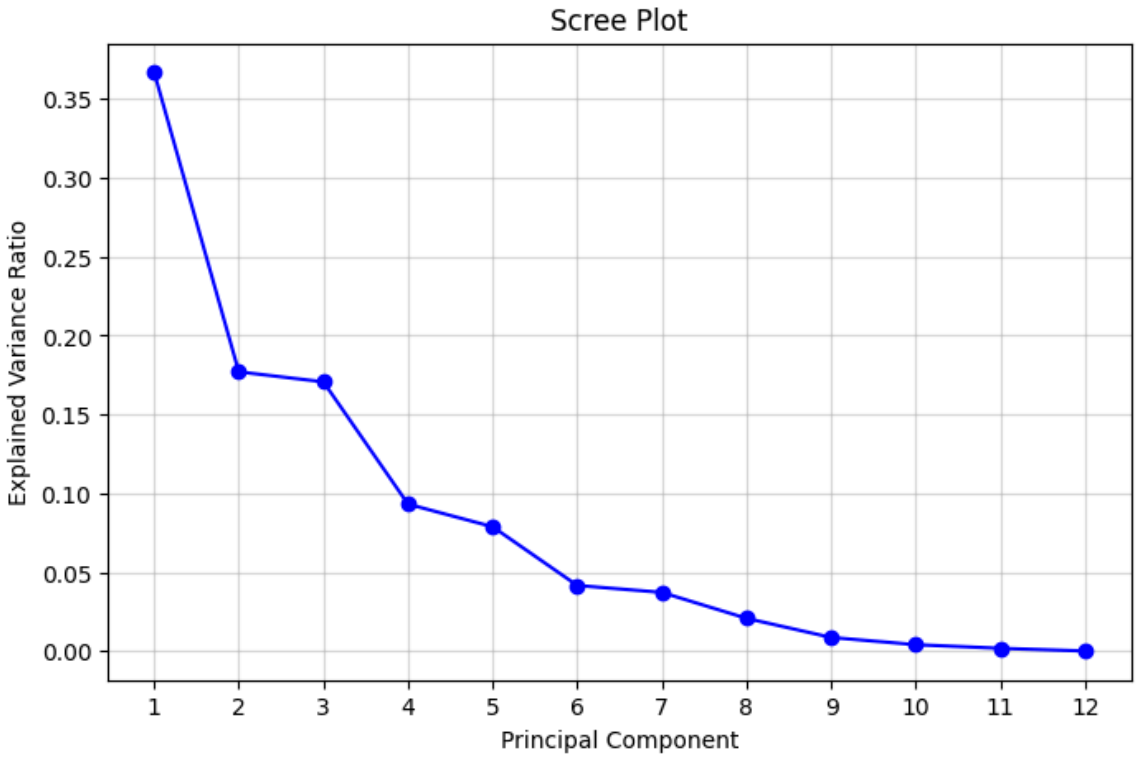
## 7. Principal Component Analysis (PCA)

The scree plot demonstrates a clear "elbow" till the 6th principal component (PC), explaining the majority of variance in the data:

- PC1 contributes ~35% of the variance.
- Cumulatively, the first 6 PCs explain a significant proportion of the data variance which is above 92%.

### 7.1. Feature Reduction:

Retaining 6 PCs optimizes computational efficiency without significant information loss



## 8. Pairwise scatterplot matrix visualization


Distinct clusters are visible, particularly between F_REAL and M_REAL, indicating that the PCA effectively captures significant variance in the data. However, there is noticeable overlap between AI-generated data (F_AI and M_AI) and real-world data (F_REAL and M_REAL), particularly in higher-order components like PC3 through PC6. The diagonal density plots show that real data groups (F_REAL, M_REAL) tend to have narrower distributions, suggesting greater consistency compared to broader spreads seen in the AI generated groups.

PC1 and PC2 appear to contribute most to separability, as they exhibit clearer distinctions between groups, while higher-order PCs add less discriminatory power. Although the PCA transformation provides a reasonable representation of the dataset's variance, the overlaps suggest that AI-generated data mimic patterns of real data to some extent.

## 9. Source Code: -

https://colab.research.google.com/drive/1eTii3rWVMey6i4jALlLkaHUTFsFgp56S#scrollTo=zU6p2IGknuc0

## 10. Bibliography: -


- The Sun, 2024. Meta reveals new AI weapon to defeat dangerous 'voice clone' deepfakes using hidden signals that your ears can't hear. [online] Available at: https://www.thesun.co.uk/tech/28606697/meta-audioseal-ai-deepfakes-release
- Wired, 2024. Deepfakes Are Evolving. This Company Wants to Catch Them All. [online] Available at: https://www.wired.com/story/deepfake-detection-get-real labs
- S. Maarek, "Aws ai practitioner certified," n.d. [Online]. Available: https://www.udemy.com/course/aws-ai-practitioner-certified/?couponCode=ST19MT121224