WaveTruth

Cracking Deepfakes with Deep Learning

What is WaveTruth

WaveTruth is an advanced web application designed to analyze audio recordings and determine:

✅ AI vs. Real – Detects whether a voice is generated by artificial intelligence or a real human.
✅ Gender Identification – Classifies the speaker as male or female based on voice characteristics.

🔍 **How It Works:**
Powered by a Feedforward Neural Network trained through transfer learning, WaveTruth ensures high-accuracy detection by using deep learning techniques in audio analysis.

🚀 Bringing Transparency to the Age of AI Voice Generation!

Scope and Audience

Scope

WaveTruth is designed to classify audio clips as either AI-generated or real human voices by analyzing key audio features such as:

🎵 Pitch Variability – Detects unnatural pitch fluctuations

🔊 Spectral Patterns – Identifies AI-generated sound signatures

🗣️ Prosody Analysis – Examines rhythm, tone, and speech dynamics

To enhance accuracy, the model is trained on a diverse dataset of AI and human voices, including varied accents, tones, pitch, and speech speed.

Audience

WaveTruth serves a broad range of users, including:

👨‍💻 Content Moderators – Detects AI-generated voices in media

🛡️ Security & Fraud Prevention Experts – Identifies synthetic speech in scams and deepfakes

🔬 AI & Voice Recognition Researchers – Advances studies in audio authentication

🚀 Bringing Trust & Transparency to Voice AI!

Background Review

Existing Approaches

Several companies have developed advanced solutions for detecting AI-generated voices by utilizing cutting-edge machine learning and deep learning techniques. For example, Evenlabs uses a variety of AI models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to detect synthetic speech. This approach allows Evenlabs to accurately identify synthetic voices, but the system is limited by access restrictions, high costs, and challenges with lower-quality or heavily edited audio. PlayHT, on the other hand, combines neural networks, including Generative Adversarial Networks (GANs) and RNNs, for both voice synthesis and detection. While PlayHT provides high-quality detection and synthesis, it also faces limitations such as potential issues with noisy audio and non-transparent pricing structures. Resemblyzer leverages wav2vec 2.0 embeddings for speaker verification and similarity detection, excelling at distinguishing specific speakers based on their voice features. However, its limitations include struggles with regional accents and languages with fewer resources.

While these companies offer innovative solutions for detecting synthetic voices, they each face unique challenges in terms of generalization across diverse audio environments and the evolving sophistication of AI voice generation techniques. As the field advances, there is a growing need for models that integrate multiple audio patterns and improve detection accuracy, which is the primary goal of this project.

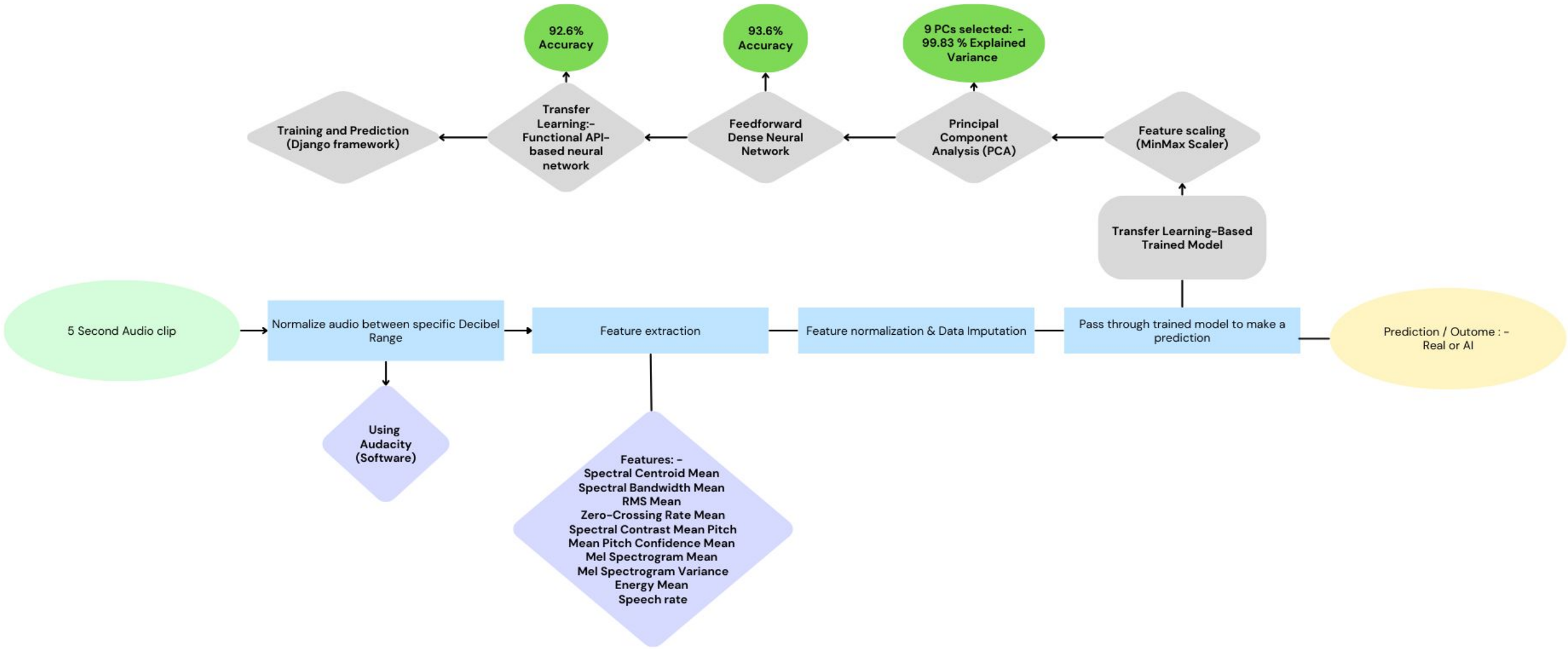
Related Literature

The exploration of voice synthesis and detection has evolved significantly over the years, with early work focusing on techniques like Mel Frequency Cepstral Coefficients (MFCCs). Recent research has increasingly turned to deep learning models such as CNNs and RNNs to classify and differentiate between human and synthetic voices. However, even these models often struggle to accurately identify new or previously unseen synthetic voices. Transformer-based architectures, which enhance context-awareness for processing audio data, have recently been introduced and show promise in improving detection accuracy. Building on this body of work, this project proposes a hybrid model that combines CNNs and transformers to better capture both temporal and spatial audio patterns, offering an improved approach to synthetic voice detection.

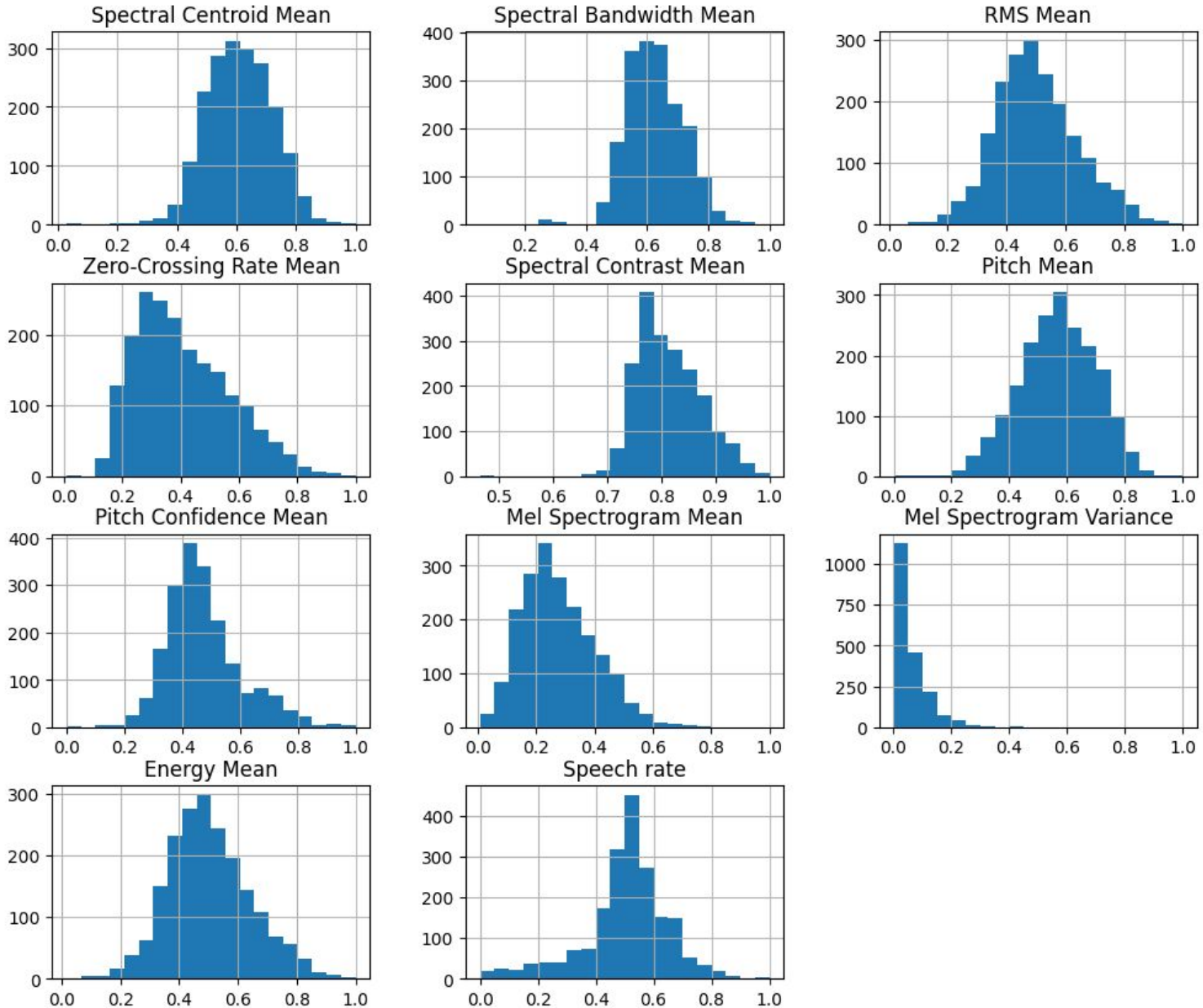
WaveTruth vs. Other Similar AI Detectors: A Comparative Analysis

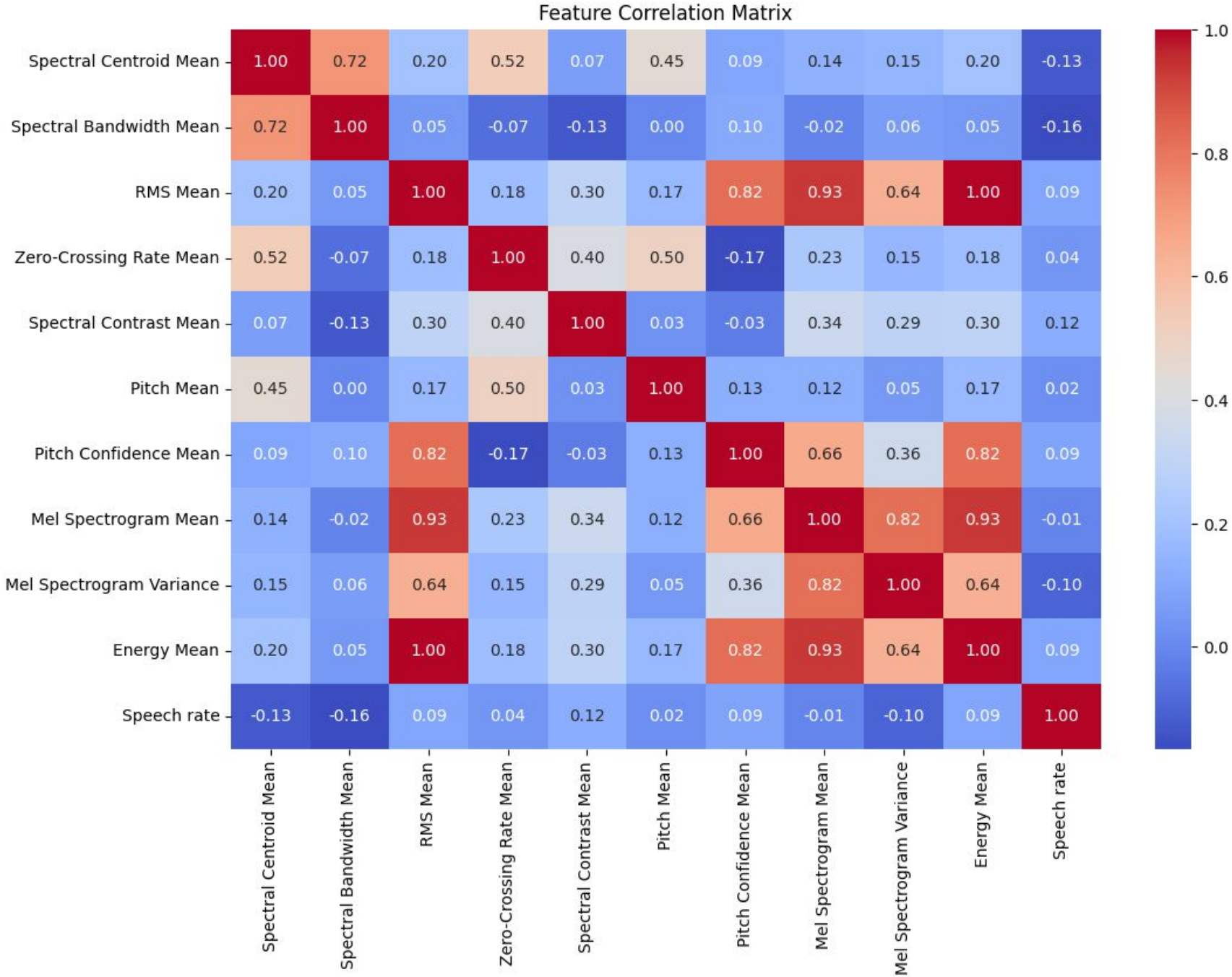
Feature	**Evenlabs**	**PlayHT**	**Resemblyzer**	***WaveTruth***
Detection Model/Algorithm	Uses a variety of AI models for synthetic speech detection, including deep learning techniques like CNNs and RNNs	Uses a combination of neural networks (including deep learning models like GANs and RNNs) for voice synthesis and detection	wav2vec 2.0 embeddings for speaker verification and similarity detection	*Feed-forward neural networks with transfer learning for gender classification*
Drawbacks	- Limited access; enterprise-focused. - Can struggle with lower-quality or heavily edited audio. - Requires commercial licensing.	- High cost for high-volume use. - May struggle with noisy or low-quality audio. - Pricing based on usage, not always transparent.	- False positives for similar-sounding speakers. - Struggles with regional accents and low-resource languages. - Limited access, mainly for research or enterprise use.	- *Limited by dataset size for underrepresented voices, though transfer learning helps mitigate this.* *- May struggle with synthetic-sounding voices.* *- Accessible via a web app only currently*
Strengths	- High accuracy in synthetic speech detection. - Well-suited for enterprise-level use and security applications.	- High-quality synthetic speech generation and detection. - Great for personalized voice creation.	- Efficient at speaker verification and comparison. - Effective for distinguishing specific speakers based on audio features.	- *Focuses on gender classification with a specialized and accessible approach.* - *Uses transfer learning to improve generalization on limited datasets.* - *Free to use*
Goal/Focus	To provide cutting-edge solutions for detecting synthetic voices and deepfakes in audio.	To offer high-quality text-to-speech synthesis and detection tools, including for AI-generated voices.	To verify and differentiate speakers based on their voice characteristics.	*To provide solutions for detecting AI audio and make gender-based voice classification accessible and reliable.*

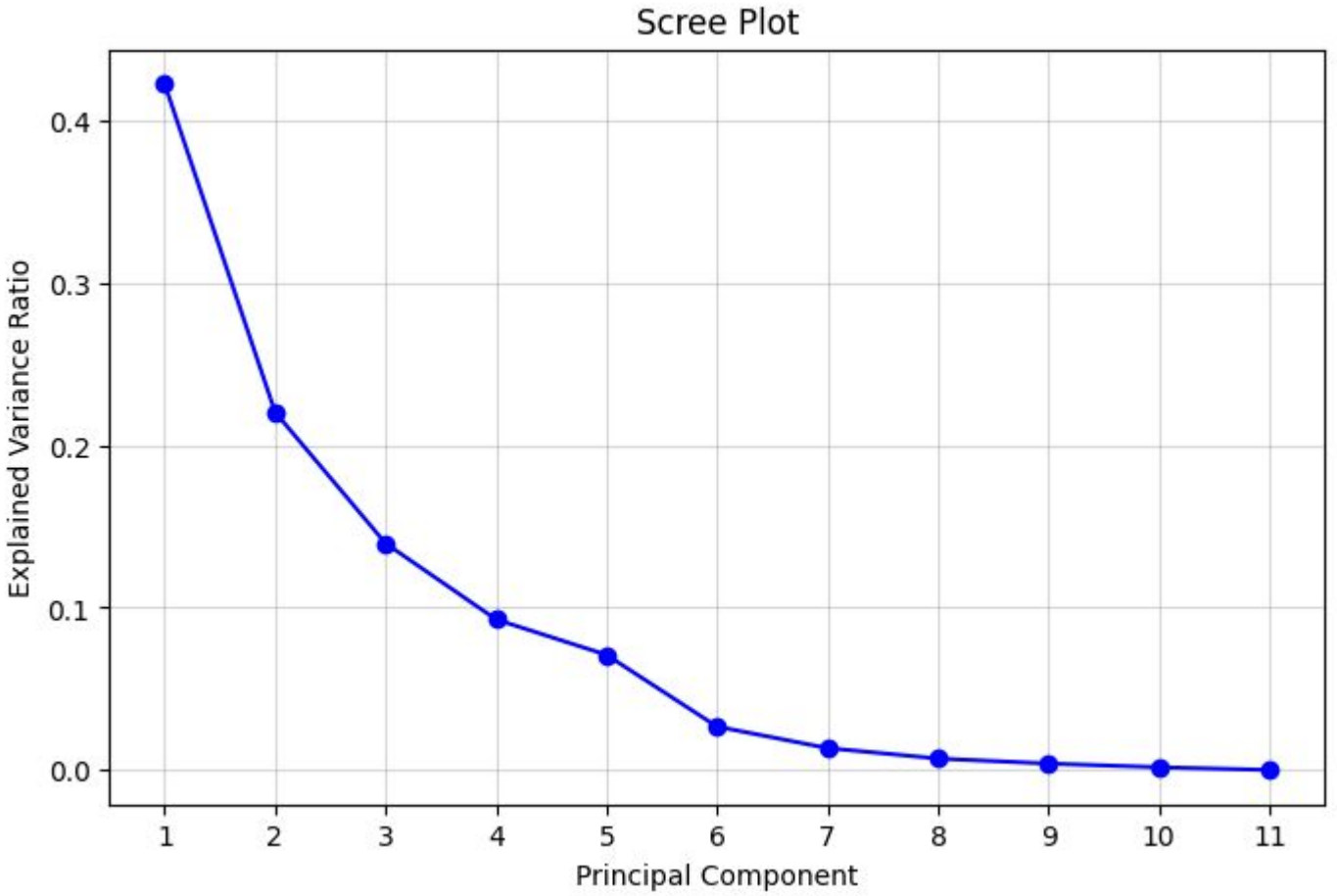
Pipeline: - Behind the scenes of how the WaveTruth makes its predictions



Feature Distributions

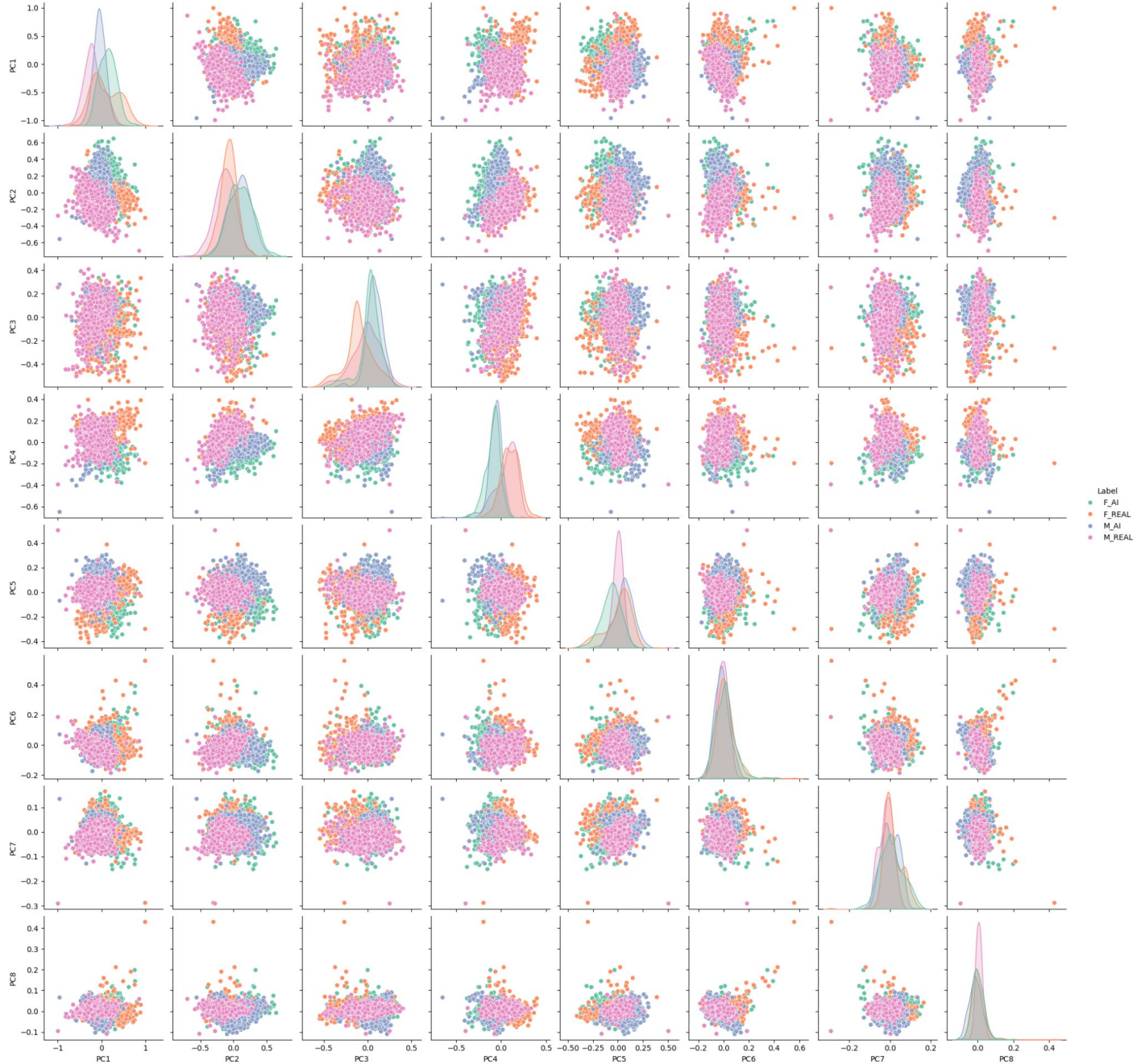


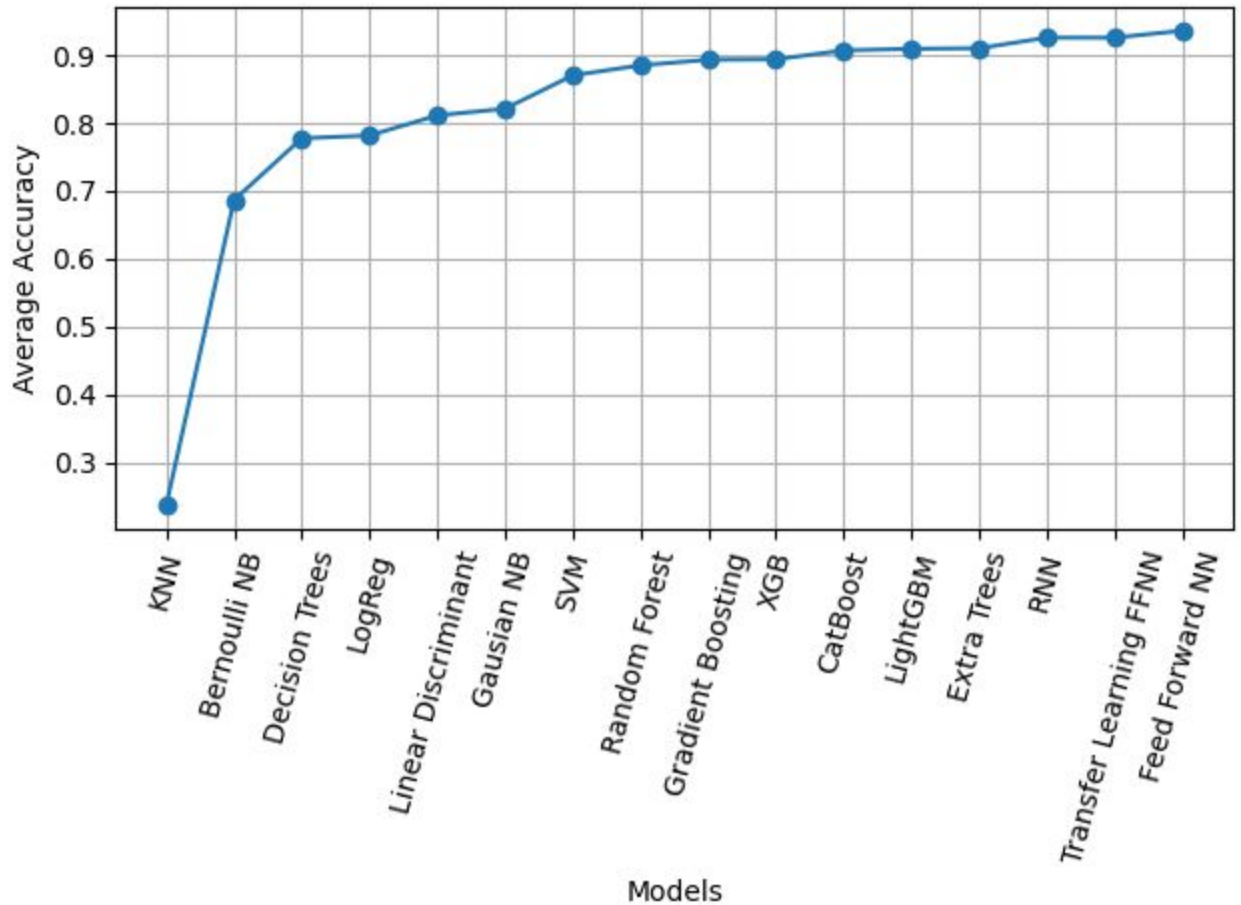

Feature Correlation Matrix
Spectral Centroid Mean 1.00 0.72 0.20 0.52 0.07 0.45 0.09 0.14 0.15 0.20 -0.13
Spectral Bandwidth Mean 0.72 1.00 0.05 -0.07 -0.13 0.00 0.10 -0.02 0.06 0.05 -0.16
RMS Mean 0.20 0.05 1.00 0.18 0.30 0.17 0.82 0.93 0.64 1.00 0.09
Zero-Crossing Rate Mean 0.52 -0.07 0.18 1.00 0.40 0.50 -0.17 0.23 0.15 0.18 0.04
Spectral Contrast Mean 0.07 -0.13 0.30 0.40 1.00 0.03 -0.03 0.34 0.29 0.30 0.12
Pitch Mean 0.45 0.00 0.17 0.50 0.03 1.00 0.13 0.12 0.05 0.17 0.02
Pitch Confidence Mean 0.09 0.10 0.82 -0.17 -0.03 0.13 1.00 0.66 0.36 0.82 0.09
Mel Spectrogram Mean 0.14 -0.02 0.93 0.23 0.34 0.12 0.66 1.00 0.82 0.93 -0.01
Mel Spectrogram Variance 0.15 0.06 0.64 0.15 0.29 0.05 0.36 0.82 1.00 0.64 -0.10
Energy Mean 0.20 0.05 1.00 0.18 0.30 0.17 0.82 0.93 0.64 1.00 0.09
Speech rate -0.13 -0.16 0.09 0.04 0.12 0.02 0.09 -0.01 -0.10 0.09 1.00
Spectral Centroid Mean
Spectral Bandwidth Mean
RMS Mean
Zero-Crossing Rate Mean
Spectral Contrast Mean
Pitch Mean
Pitch Confidence Mean
Mel Spectrogram Mean
Mel Spectrogram Variance
Energy Mean
Speech rate
1.0
0.8
0.6
0.4
0.2
0.0

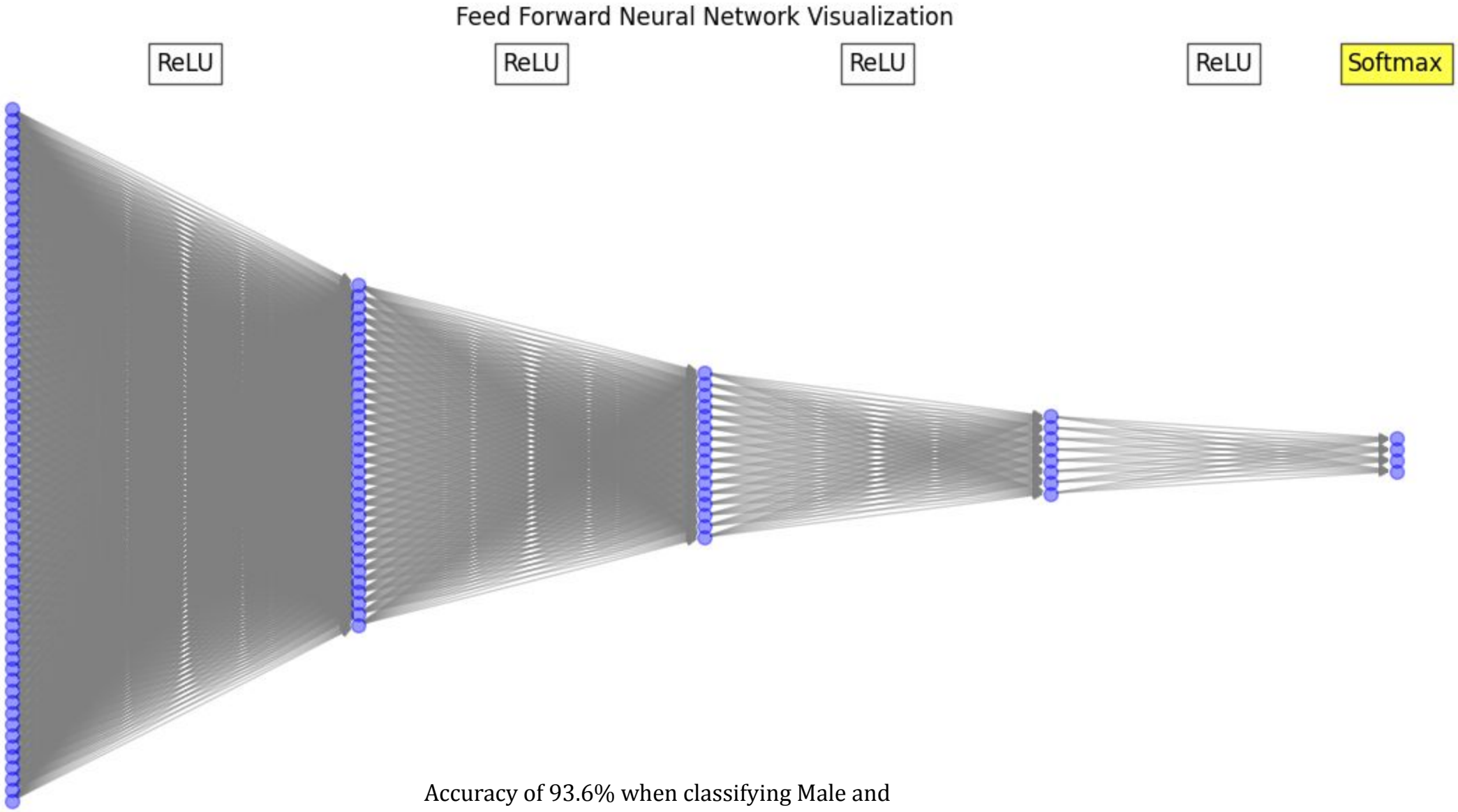


9 Principal Components Selected, 99.83 % of total explained Variance

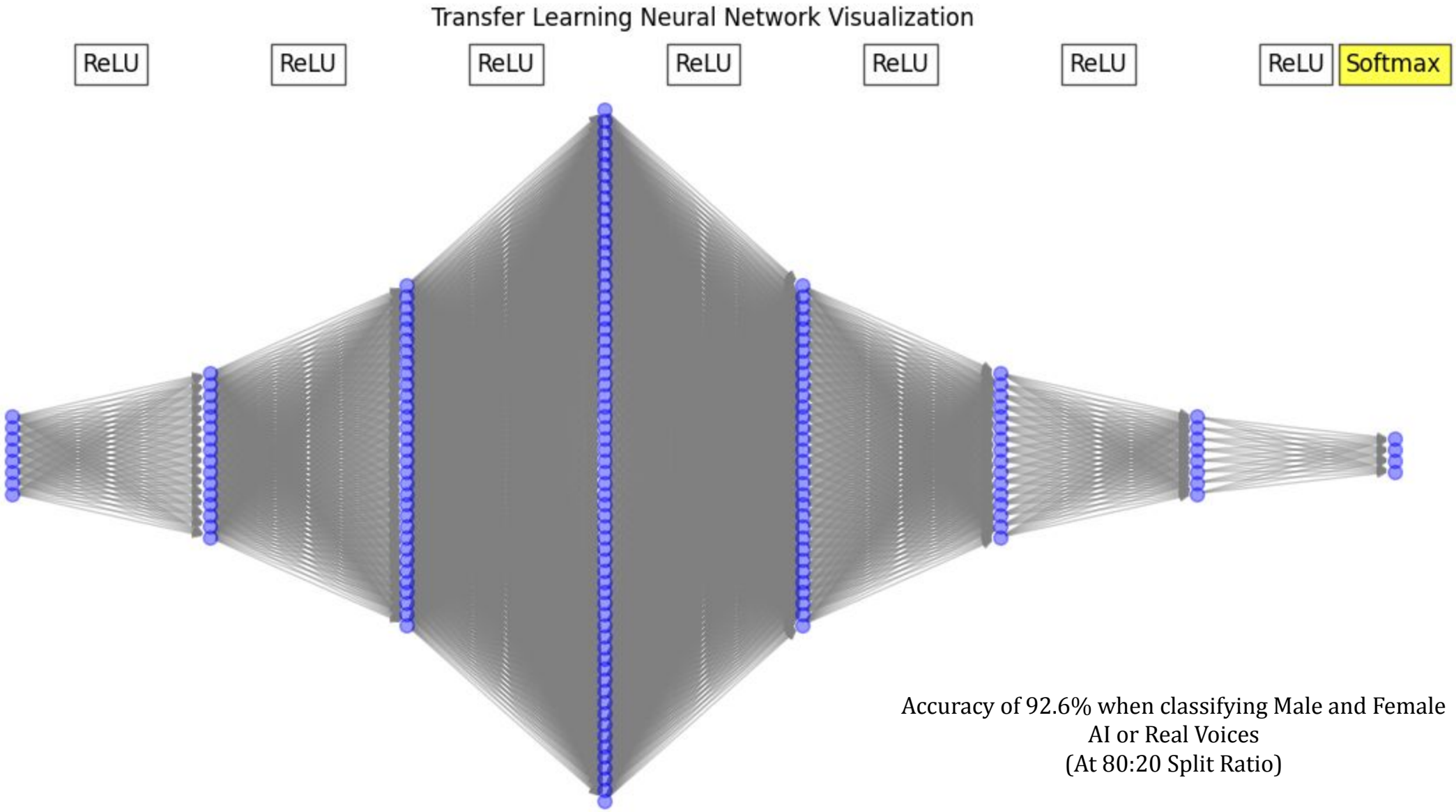
Pairwise Scatter Plots of Principal Components




Average Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
KNN
Bernoulli NB
Decision Trees
LogReg
Linear Discriminant
Gausian NB
SVM
Random Forest
Gradient Boosting
XGB
CatBoost
LightGBM
Extra Trees
RNN
Transfer Learning FFNN
Feed Forward NN
Models



Accuracy of 93.6% when classifying Male and
Female AI or Real Voices
(At 80:20 split ratio)



Accuracy of 92.6% when classifying Male and Female
AI or Real Voices
(At 80:20 Split Ratio)

UI development in progress

Django Framework