His, Hers, Hertz

Gender Classification through Audio Feature Extraction and Machine Learning

Presented by:
Jason Catacutan

Date Submitted:
March 23, 2025

Agenda

- Overview
- Methodology
- Webscraping & Feature Extraction
- Preprocessing & Data Set
- Exploratory Data Analysis
- Model Building
- Results & Discussion
- Conclusion

Overview

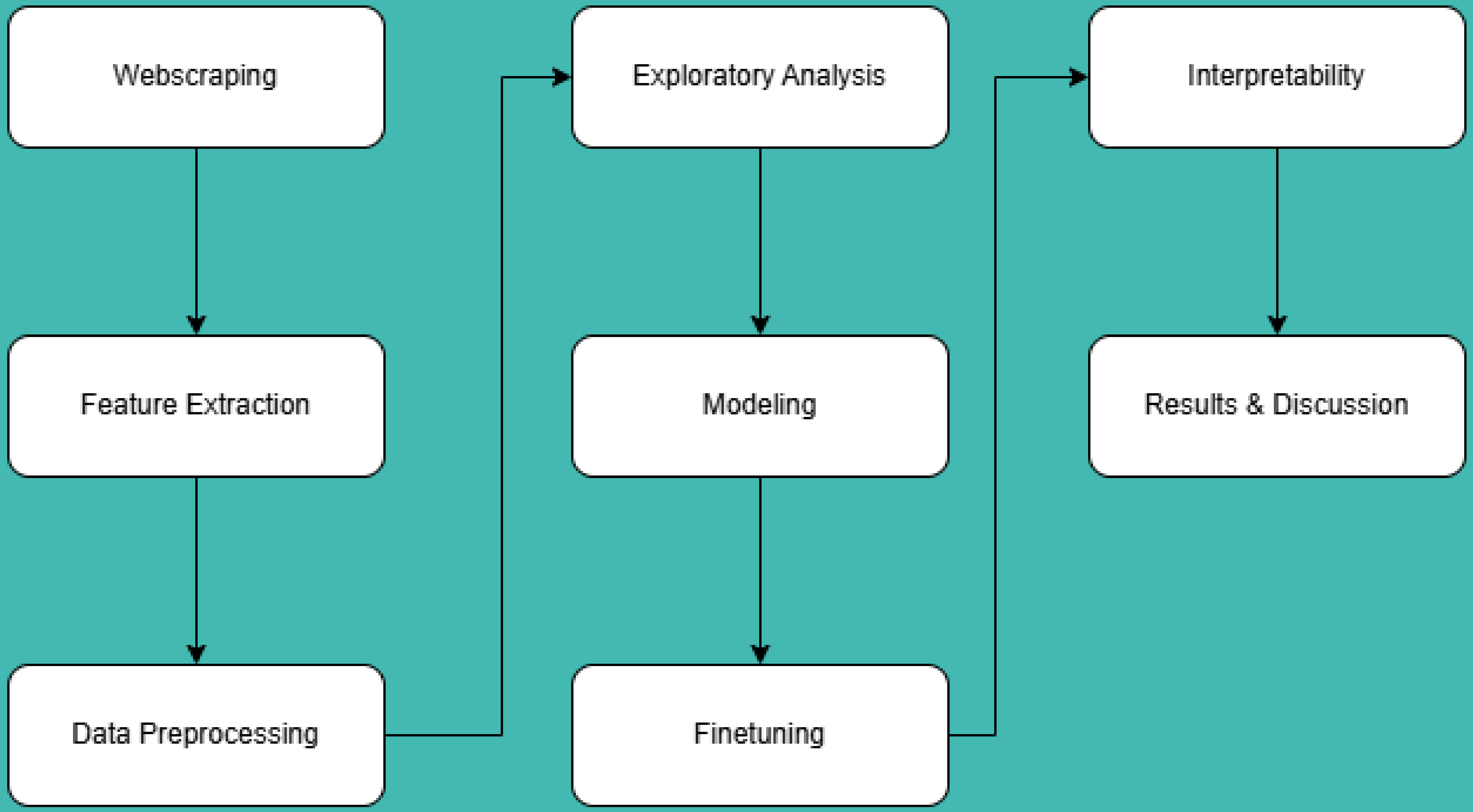
Objective

To extract, investigate, and analyze audio data from English-speaking male and female samples, aiming to develop a machine learning model capable of predicting gender based on distinctive vocal features.

Data Provided

The raw dataset provided by VoxForge consists of compressed TGZ files containing .wav audio files along with other related materials for each sample. When fully decompressed, the dataset size is expected to be approximately 12.5 GB.

Methodology



Webscraping & Feature Extraction

01 Raw Data

Create a Python script to automate scraping and extracting TGZ files from VoxForge.

02 Filtering

Filter raw data files by discarding those with >90% noise outside the human vocal range.

03 Statistics

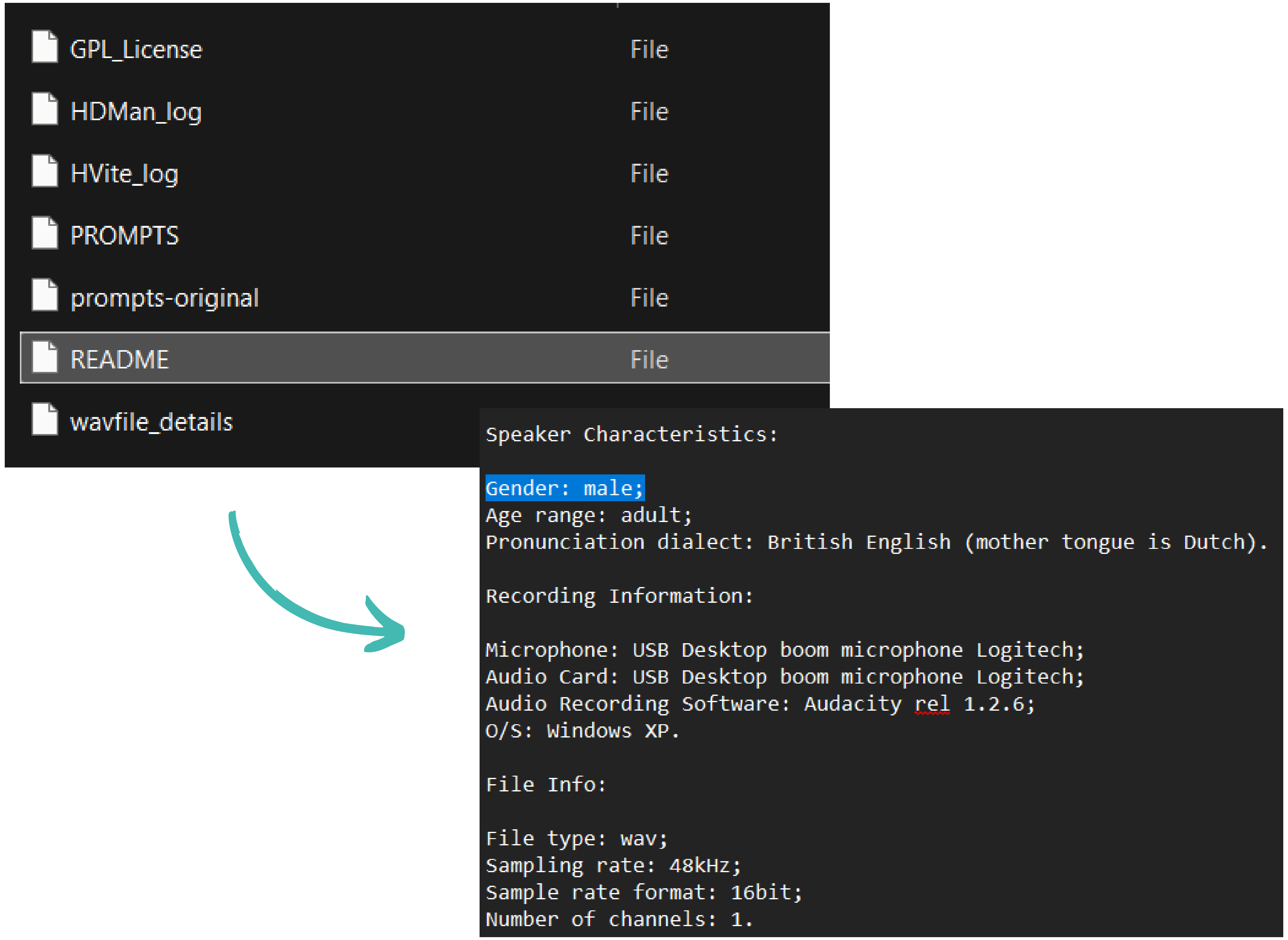
Extract statistical features using Numpy and Scipy stat functions.

04 Librosa

Use Librosa to compute the FFT spectrum of audio data

Preprocessing & Data Set

Acquiring the Target Variable



The target variables (Male or Female) were located in a README file within the TGZ archive. These were automatically extracted using Regex and merged with the final dataset.

Remapping Labels

Age Range

```
merged_df["age_range"] = (
    merged_df["age_range"]
    .str.lower()
    .str.replace(";", "", regex=False)
    .str.strip()
    .replace({
        "erwachsener": "adult",
        "adulto": "adult",
        "adulte": "adult",
        "adult (born in 1983)": "adult",
        "[adult]": "adult",
        "[adult]": "adult",
        "youth;": "youth",
        "[youth]": "youth",
        "jeune": "youth",
        "senior;": "senior",
        "please select": "unknown",
        None: "unknown",
        "male": "unknown"
    })
    .fillna("unknown")
)
```

Gender

```
gender_map = {
    'male': 'male',
    'make': 'male',
    'männlich': 'male',
    'masculino': 'male',
    'masculin': 'male',
    'female': 'female',
    'weiblich': 'female'
}

merged_df['gender'] = merged_df['gender'].map(gender_map)

merged_df = merged_df[merged_df['gender'].isin(['male', 'female'])]
```

Certain variables, such as age range and gender, were mislabeled or used inconsistent terminology, requiring remapping for accuracy and uniformity.

Dataset

Final Count

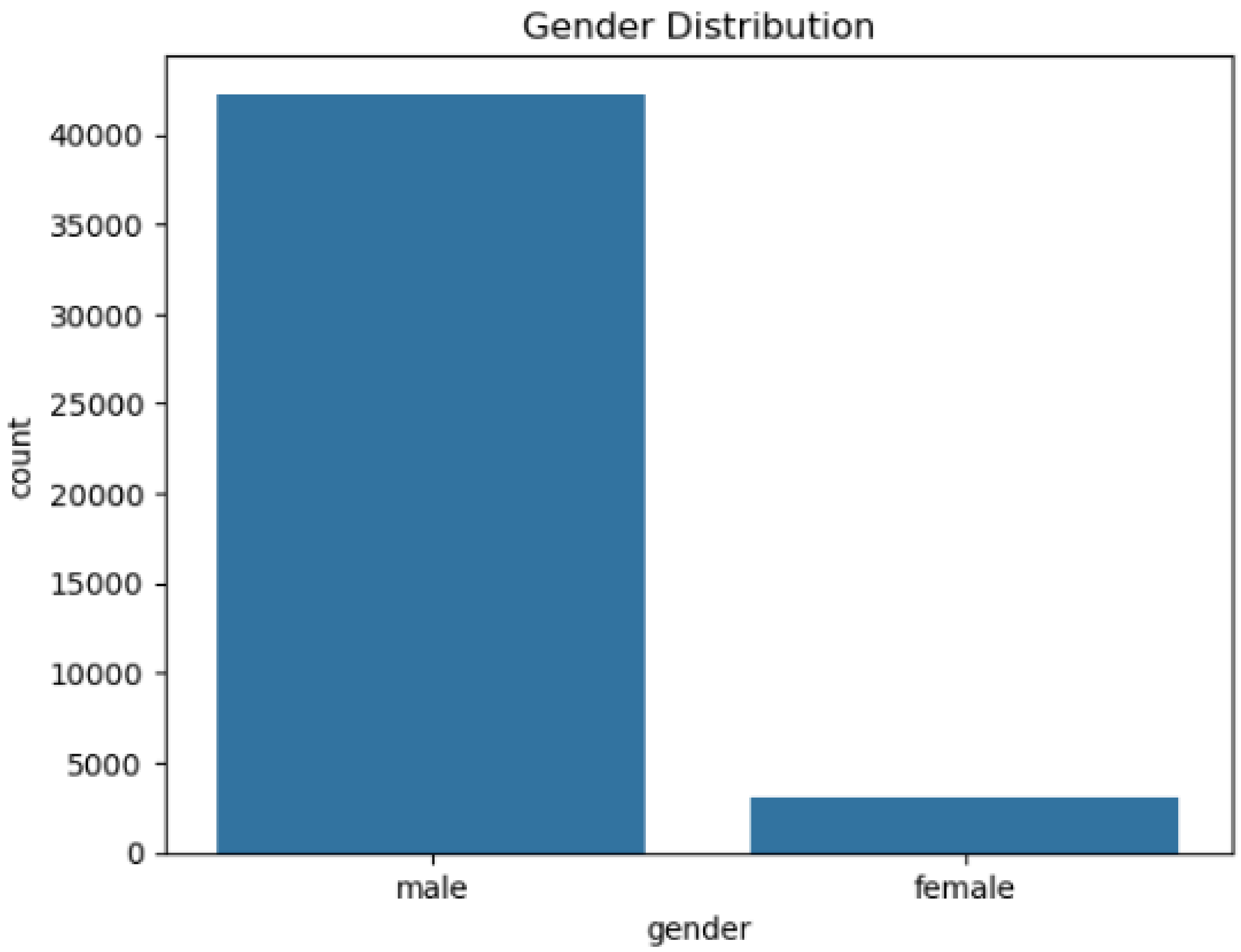
After additional filtering to ensure all samples were in English, along with the removal of missing values and duplicates, the final dataset consisted of 45,295 individual .wav files.

Features

The dataset features include: filename, mean frequency (kHz), standard deviation of frequency (kHz), median frequency (kHz), first quantile (kHz), third quantile (kHz), interquartile range (kHz), skewness, kurtosis, mode frequency (kHz), peak frequency (kHz), spectral entropy, flatness, centroid (kHz), modulation index, gender, age range.

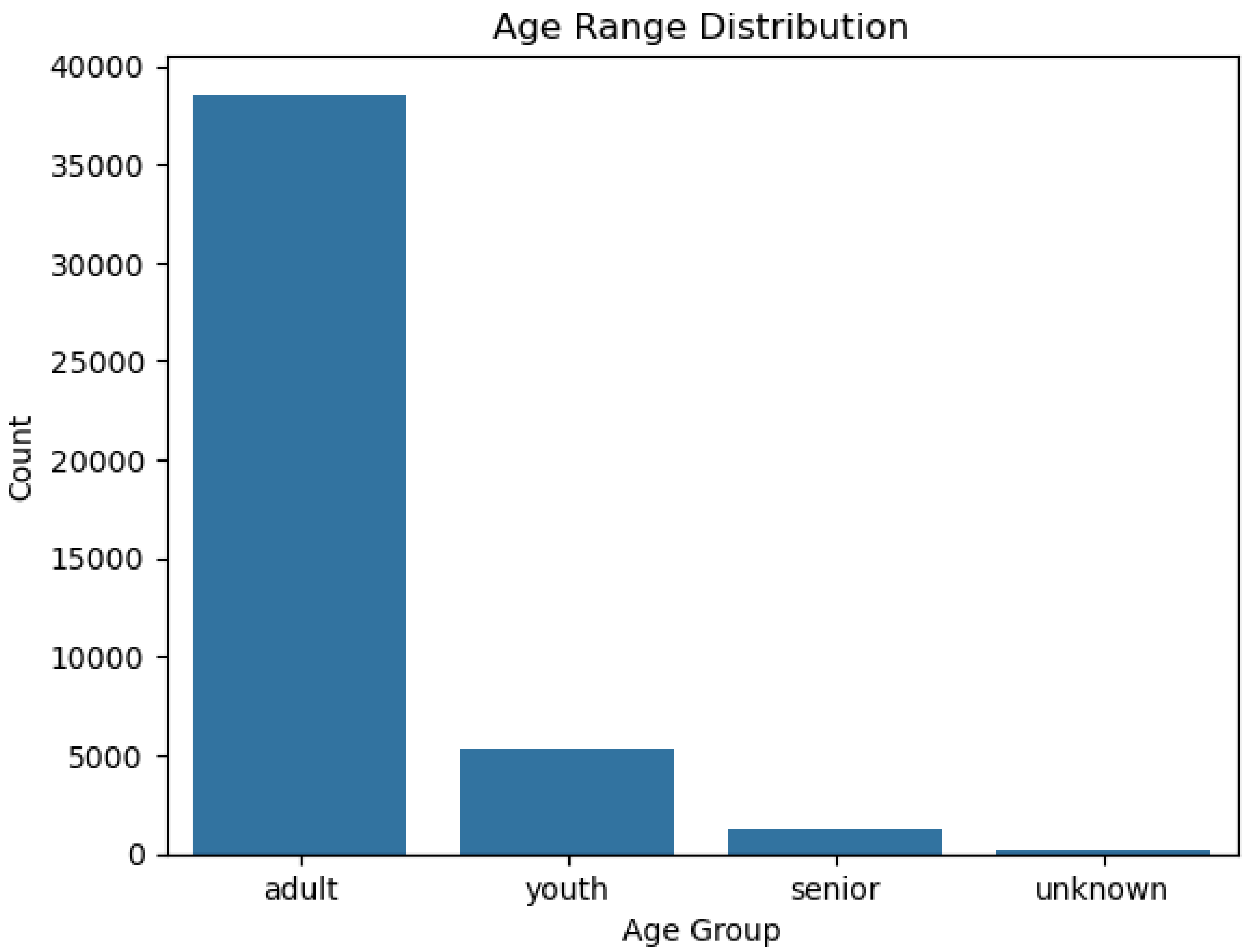
Exploratory Data Analysis

Class Imbalance - Target



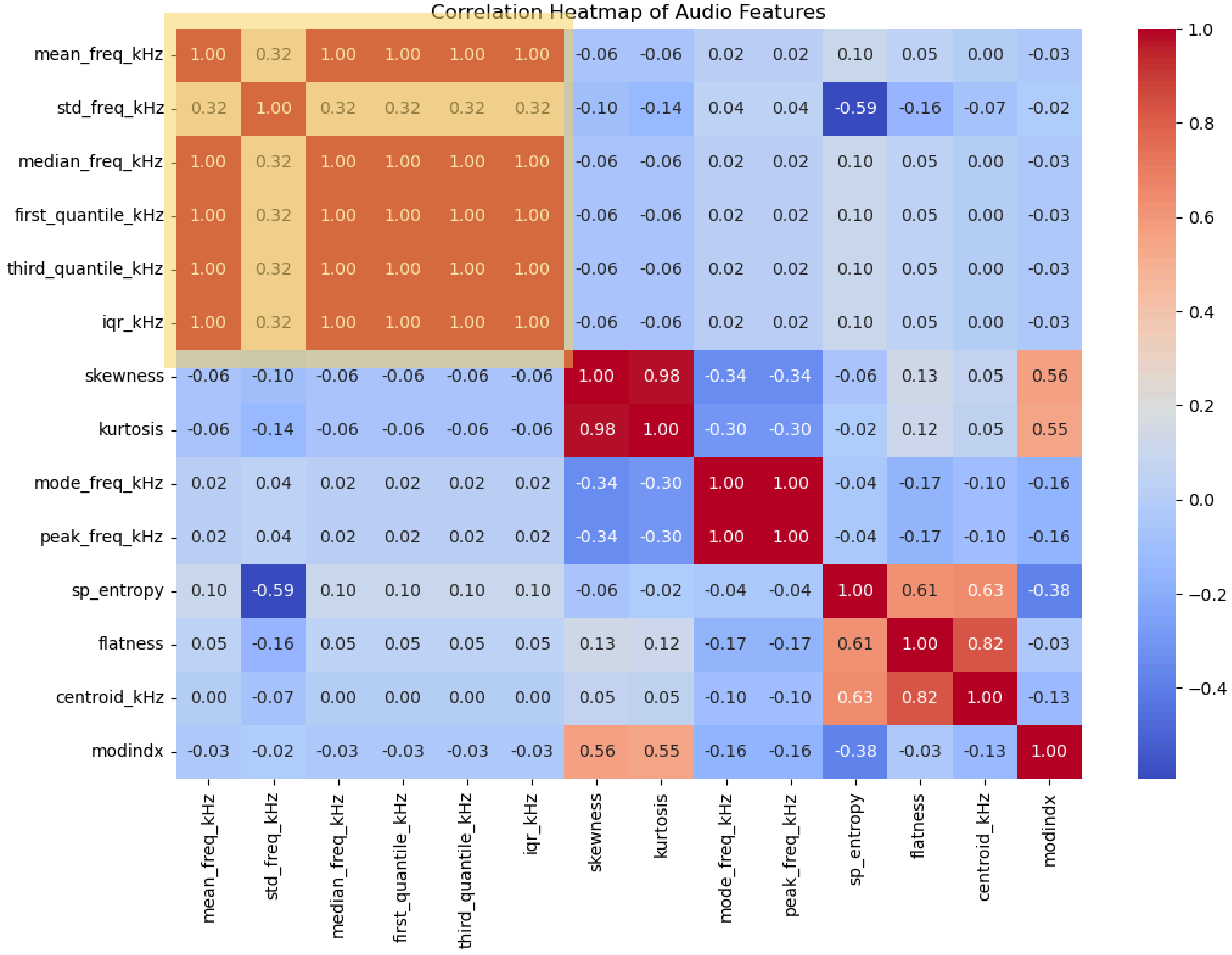
The dataset exhibits a significant class imbalance in the target variable, which could present challenges during the modeling process.

Class Imbalance - Age



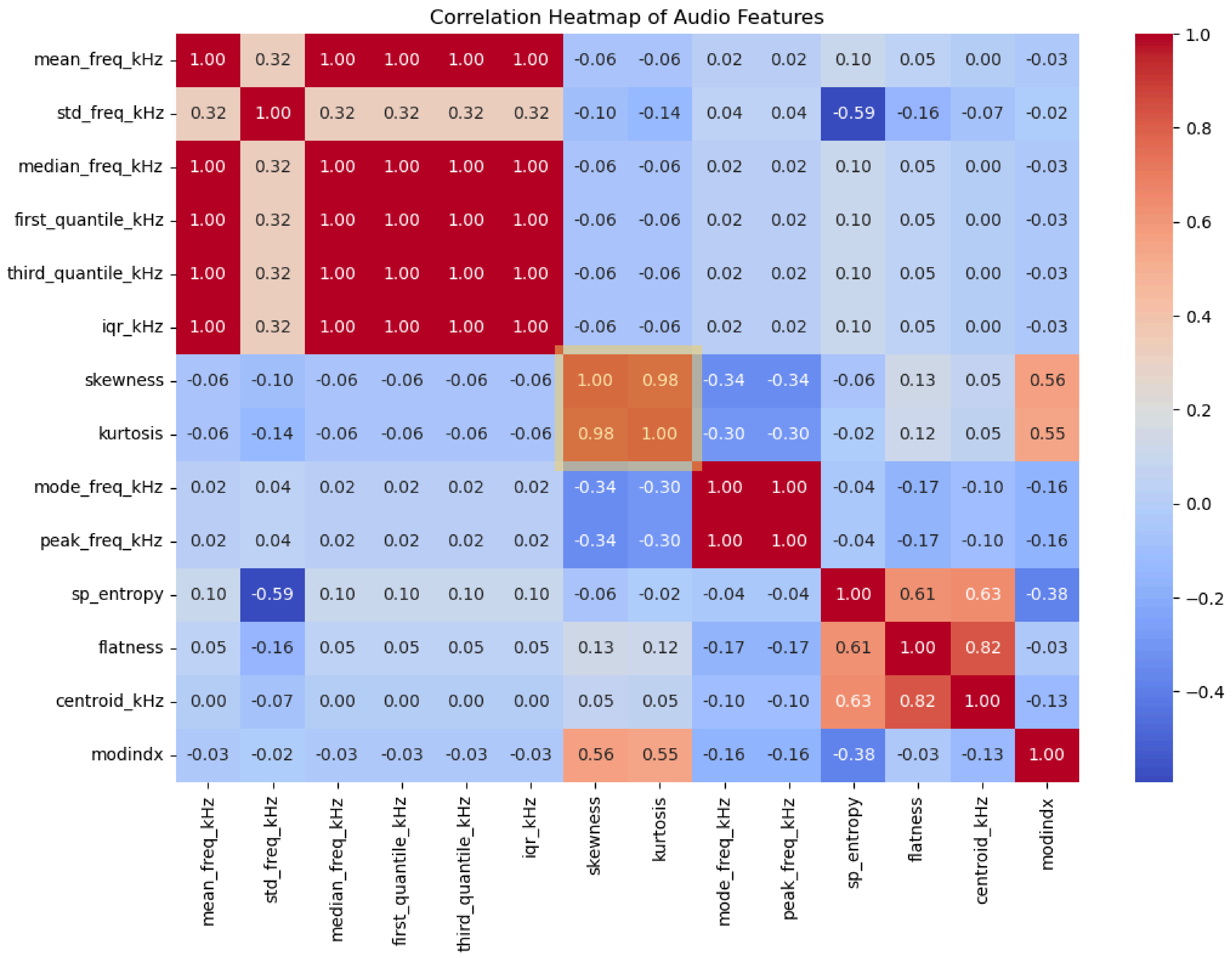
Class imbalance was also noted in the age distribution, though it was less severe compared to the target variable. This issue may be addressed at a later stage.

Correlation Matrix



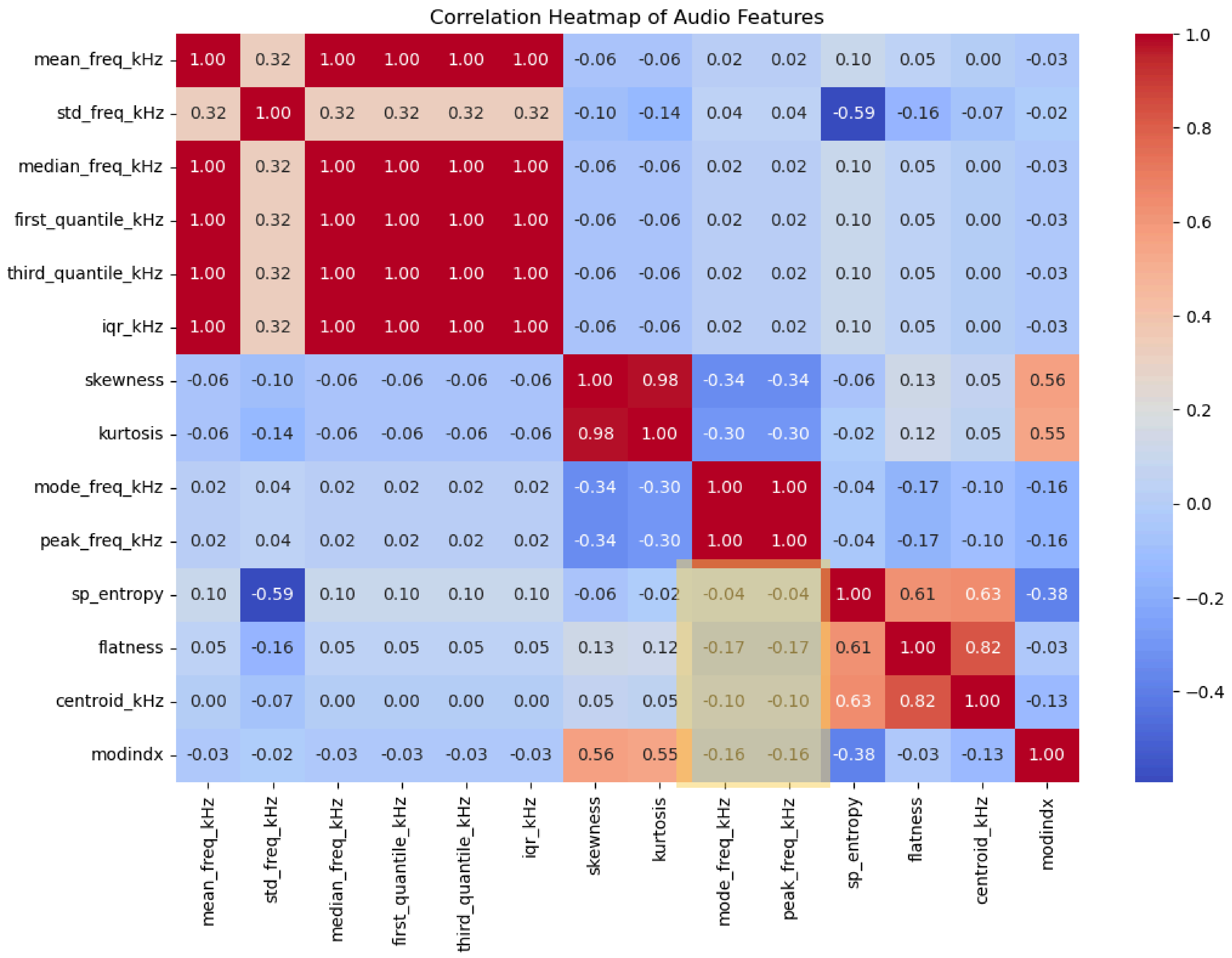
Correlation analysis revealed several highly redundant frequency-based features (e.g., mean, median, and quantiles)

Correlation Matrix



Spectral shape features like skewness and kurtosis also showed strong overlap

Correlation Matrix



Features like spectral entropy, flatness, centroid, and modulation index provided unique, low-correlated signals

Model Building

Set Up

Feature Selection

Based on insights from EDA, the features were narrowed down to: mean_freq_kHz, std_freq_kHz, skewness, kurtosis, mode_freq_kHz, sp_entropy, flatness, centroid_kHz, and modindx.

Data Split

The data was divided into Training, Validation, and Test sets to enhance robustness. To prevent data leakage, the split was performed at the file level rather than the .wav level, ensuring that samples from the same individual were confined to a single set.

Model Selection

Various models, including basic regression, kNN, tree-based approaches, and boosting methods, were utilized to determine the best fit for the task

Baseline

MLP Classifier	Precision	Recall	F1	Support
Female	0.71	0.08	0.14	329
Male	0.95	1.00	0.98	6426

With default parameters, the MLP Classifier achieved a 95.35% test accuracy. However, issues with actual performance were evident when analyzing Recall and F1 scores.

Resampling

MLP Classifier	Precision	Recall	F1	Support
Female	0.15	0.45	0.23	329
Male	0.97	0.87	0.92	6426

Various resampling techniques, including Oversampling, Undersampling, SMOTE, and ADASYN, were applied to enhance model performance.

Among the tested models, XGBoost combined with random oversampling delivered the most significant improvement in F1 score while maintaining a respectable test accuracy of 85%.

Finetuning w/ Optuna

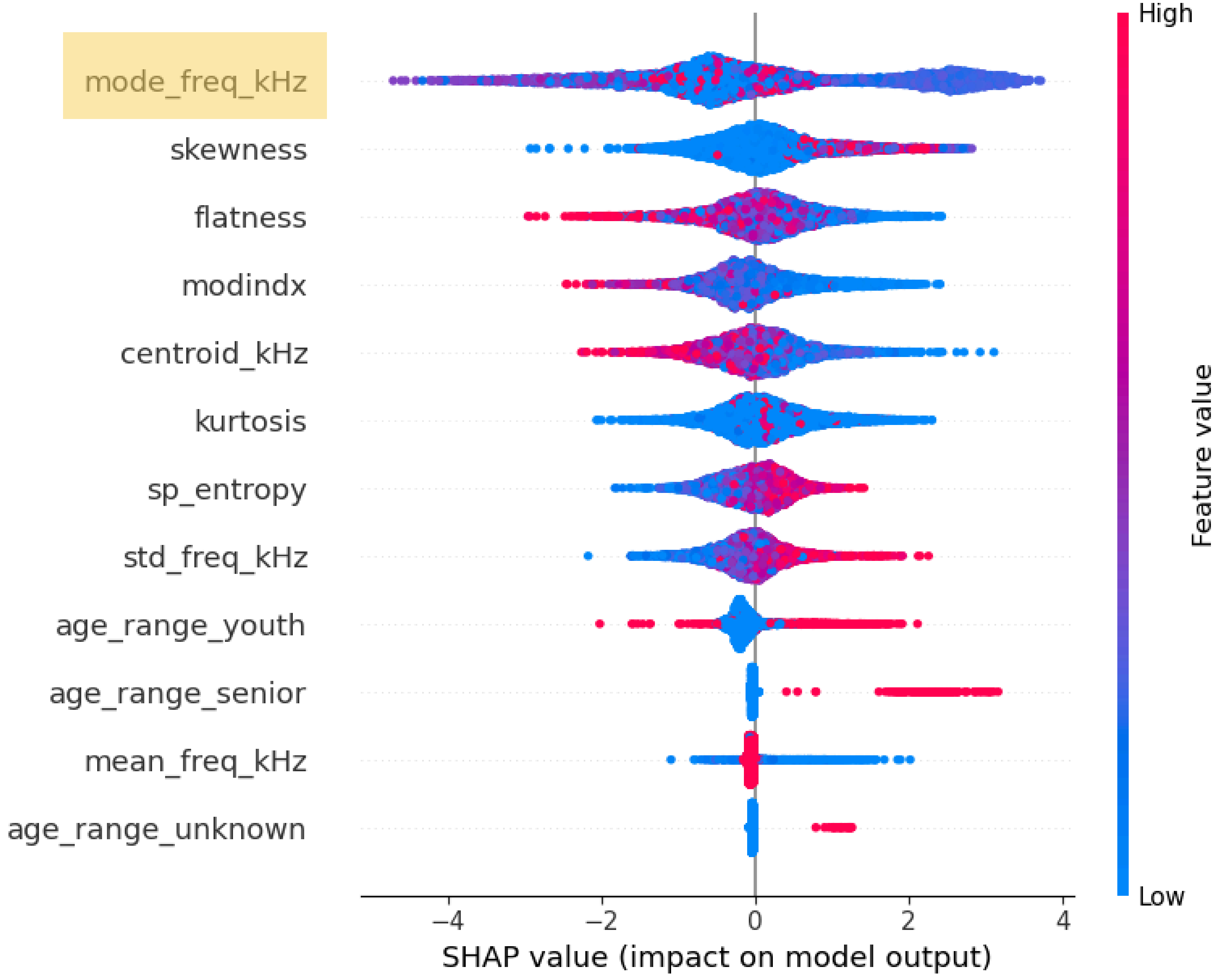
MLP Classifier	Precision	Recall	F1	Support
Female	0.33	0.30	0.32	329
Male	0.96	0.97	0.97	6426

To enhance performance, the Optuna library was utilized for hyperparameter optimization, replacing manual grid search.

This approach led to a notable improvement, achieving a test accuracy of 93.69% and significant gains in the F1 score.

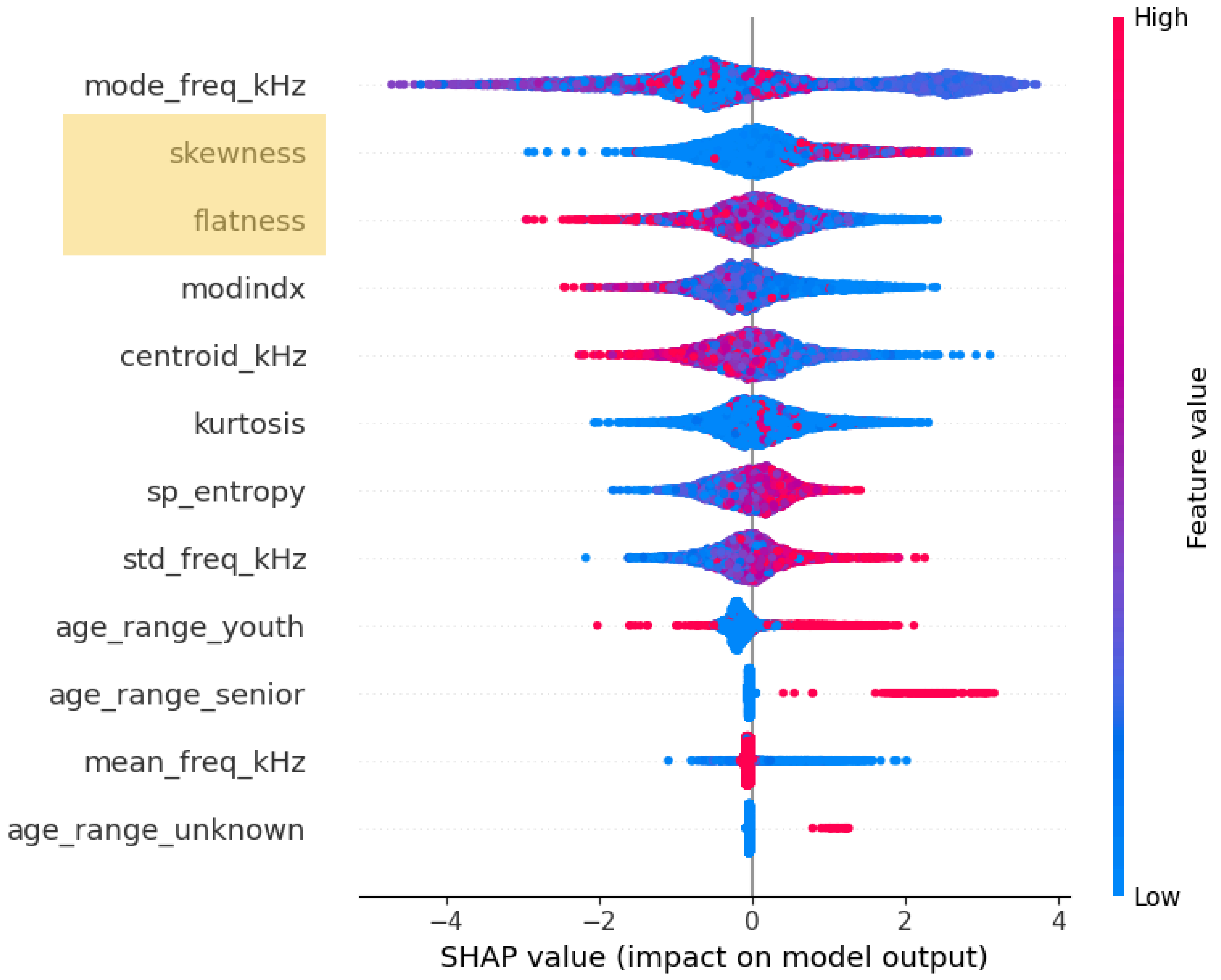
Results & Discussion

SHAP



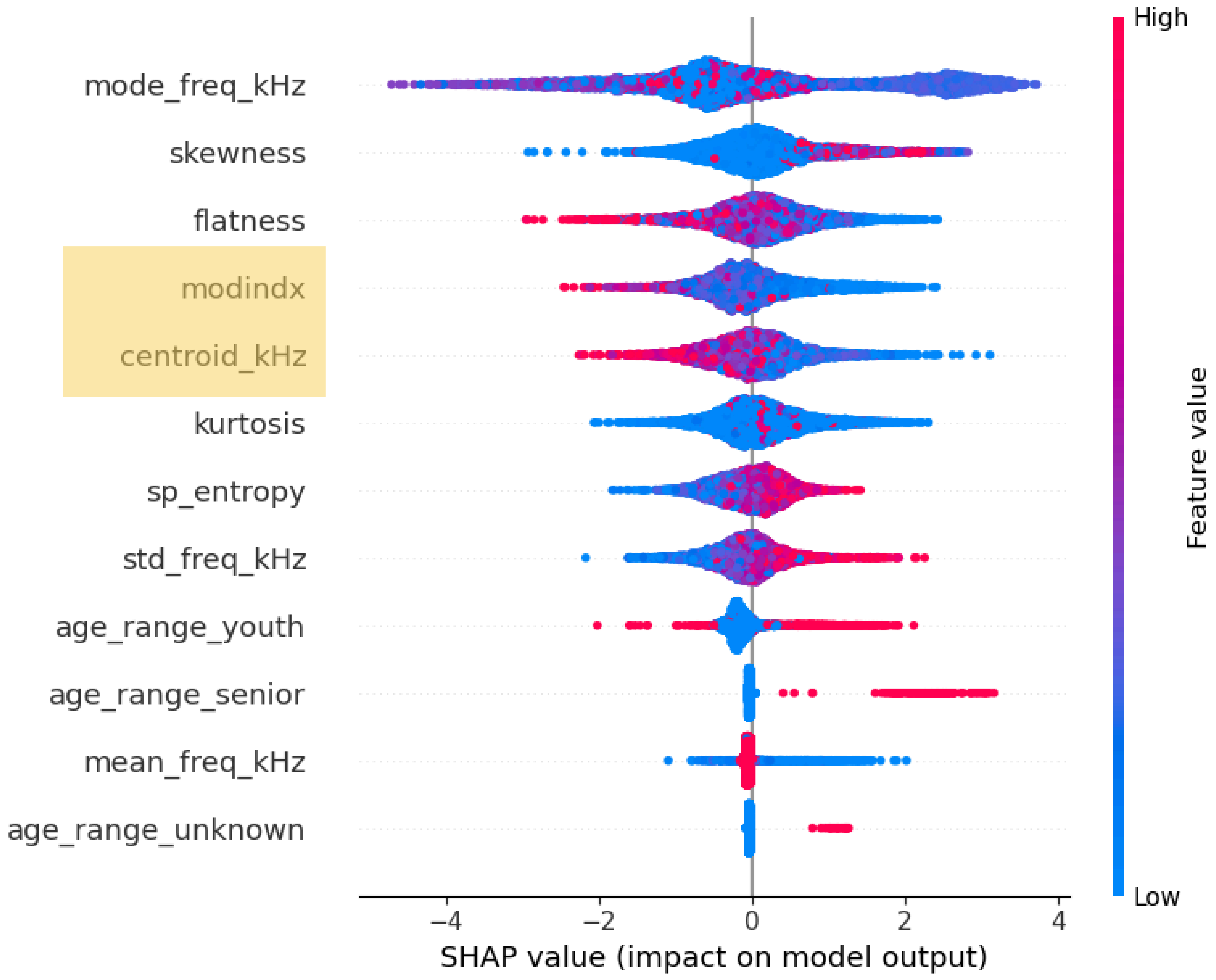
Mode_freq_kHz was identified as the most influential feature, with low values generally indicating male predictions, while high values, though less clear-cut, increased the likelihood of female classification.

SHAP



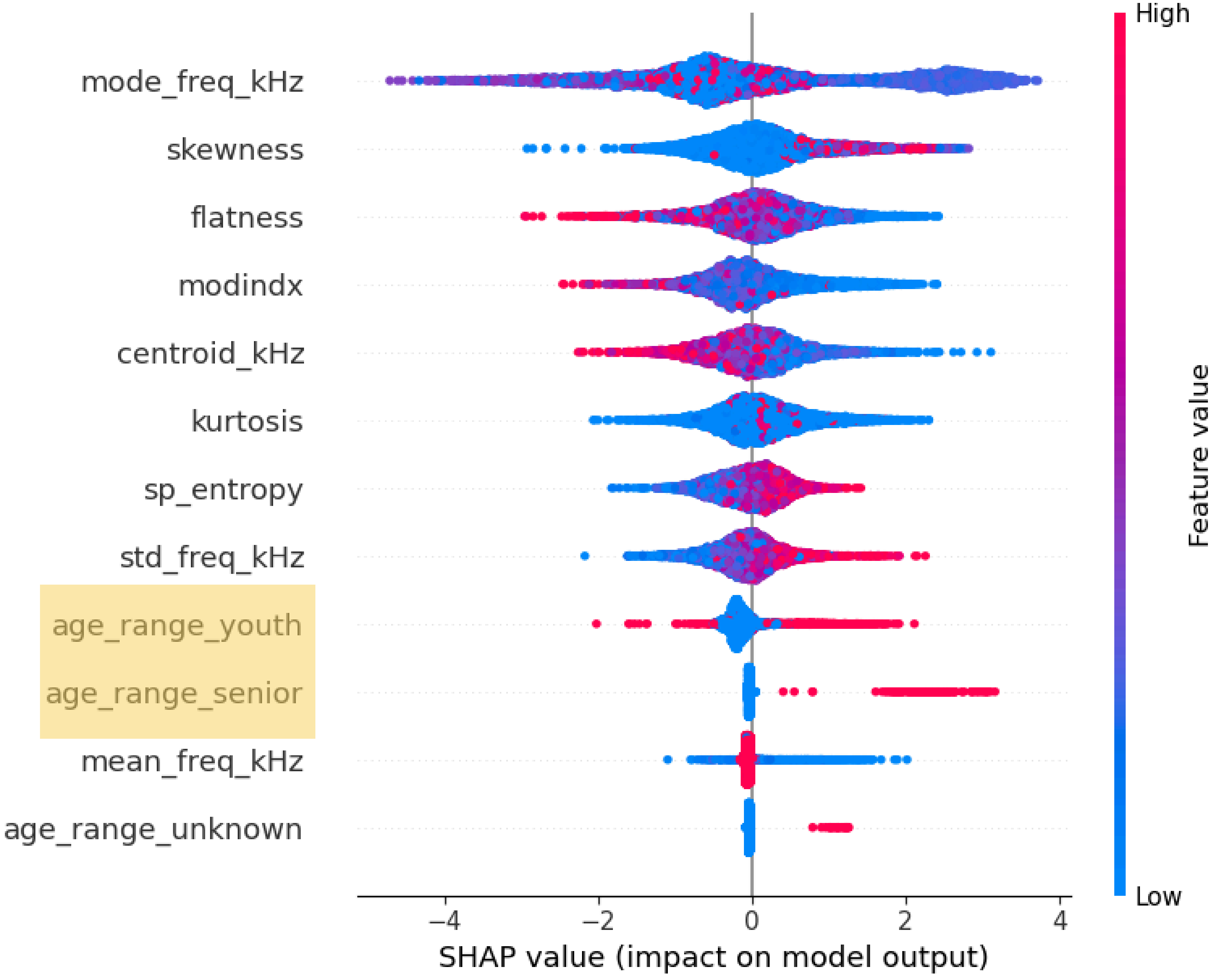
Skewness and flatness significantly characterize spectral shape, with low skewness linked to female predictions and lower flatness leaning toward male classification.

SHAP



Modindx and centroid_kHz showed decisive predictions, with high modulation index and centroid_kHz favoring female classification. .

SHAP



Age features had minimal impact, with senior speakers slightly favoring male predictions.

This can hint at the influence of aging (puberty) can affect the distinguishability of the voice

Conclusions & Recommendations

Recommendations for Model Performance

01 More Data on Females

Acquiring more data on the underrepresented class can improve model perforance and addres imbalance

02 Feature Engineering

Acquiring more features, that provide better distinction across genders can help better classify classes

03 Transformers

Using more powerful architecture, like transformers, can yield better results. However, it may require more/different data and compute power

Recommendations for Enterprise Application

01 Data Enrichment

Customer gender can be inferred from voice interactions, such as call logs, to enhance CRM data without relying on explicit survey responses.

02 Fraud Detection

Automatic gender classification can help build more robust multifactor authentication systems

03 Client Segmentation

Inferred gender can expedite segmentation tagging, whether for credit risk, health, marketing, etc.

Conclusion

Acoustic Features

Mode frequency, spectral shape (skewness, flatness), and modulation index emerged as the most important features, aligning with both prior literature and model explainability tools like SHAP.

Balancing & Finetuning

Resampling techniques, combined with optimized fine-tuning, effectively address the challenges of an imbalanced dataset while minimizing trade-offs in overall performance.

Real World Application

Gender inference from voice can support enterprise applications, particularly in scenarios where demographic data is unavailable or incomplete.

Thank You