**Gender Classification from Audio Task Report**

**1. Dataset Description**

The dataset utilized in this project consists of dev-clean speech data from LibriSpeech, comprising a development set with 'clean' speech recordings. The gender labels are incorporated into the folder names along with the speaker ID, formatted as 'speaker ID\_F/M'. To optimize computational resources, a subset of 20 speakers' speeches was retained, evenly split between 10 females and 10 males, all in FLAC format.

**2. Audio Features Analysis**

Due to challenges encountered in converting MFCCs features into an array, alternative features were extracted from the audio files, including audio length, tempo, mean fundamental frequency (F0), and standard deviation of F0 for each audio file. Notably, features such as speech rates and median F0 were omitted due to computational constraints. 'Male' is represented as 0, while 'Female' is represented as 1. Descriptive statistics and visualizations were conducted using Jamovi and Python. Additionally, Mel-spectrograms were extracted for training Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) models.

GLME analysis using Jamovi was performed on tempo, mean F0, and SD of F0 across sex groups to explore potential effects of personal speech characteristics on the results.

**2.1 Audio Length**

The total audio length for each sex is comparable (Female = 4841.89 secs, Male = 4841.41 secs), with mean lengths of 7.38 secs for female audio and 7.10 secs for male audio.

**2.2 Tempo and F0**

GLME analysis indicated that female speakers tend to speak slightly quicker than male speakers (Mean: F = 124 bpm, M = 120 bpm), with greater variance observed among male speakers (SD: F = 30.7, M = 33.1). However, no significant effect of sex on tempo was observed (F 1, 19.6 = 3.41, p = 0.08).

Male speakers exhibited lower mean F0 (mean = 153, median = 121) compared to females (mean = 234, median = 212), with higher variance among male speakers (SD = 215) than female speakers (SD = 108). The GLME analysis revealed a significant effect of sex on mean F0 (F 1, 18.4 = 13.3, p = 0.002), indicating that sex accounts for a substantial portion of the variability in mean F0. Additionally, there was a significant effect of group on SD of F0 (F 1, 18.0 = 5.26, p = 0.034), suggesting differences in within-group variance.

**2.3 Data Preprocessing**

One female speech audio was found to have missing data for mean F0 and SD of F0, attributed to speaker 6313. Given the audio's short length of 2.29 secs, its removal is not expected to significantly impact classifier performance, warranting its deletion from the dataset.

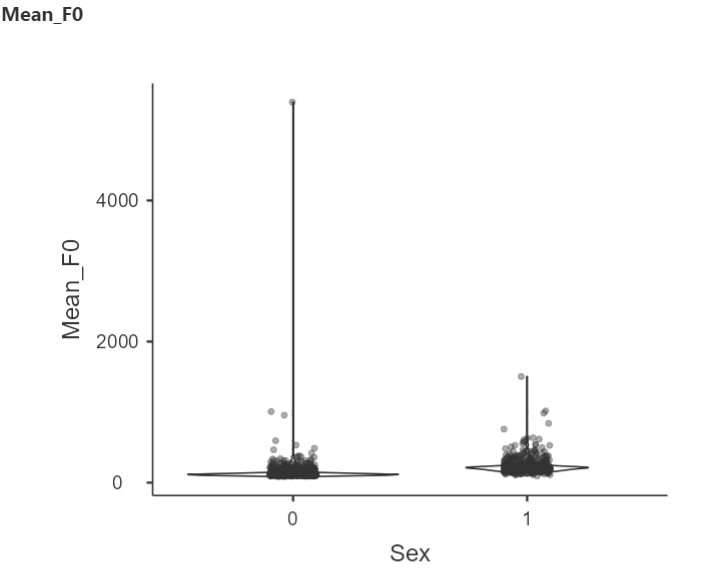


Figure 1: Violin plots of mean F0 across two sex groups with 0 representing males and 1 representing females. Jittered dots represent individual values.

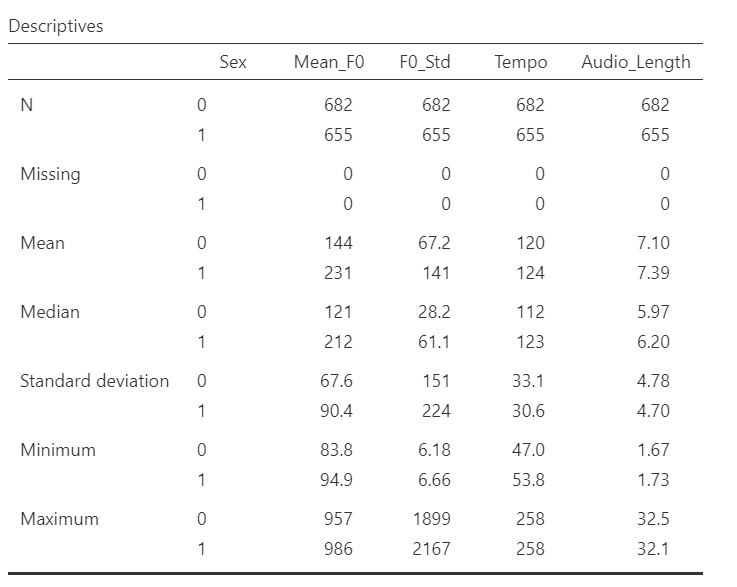


Table 1: Descriptive statistics of mean f0, SD of f0, tempo, and audio length after data preprocessing.

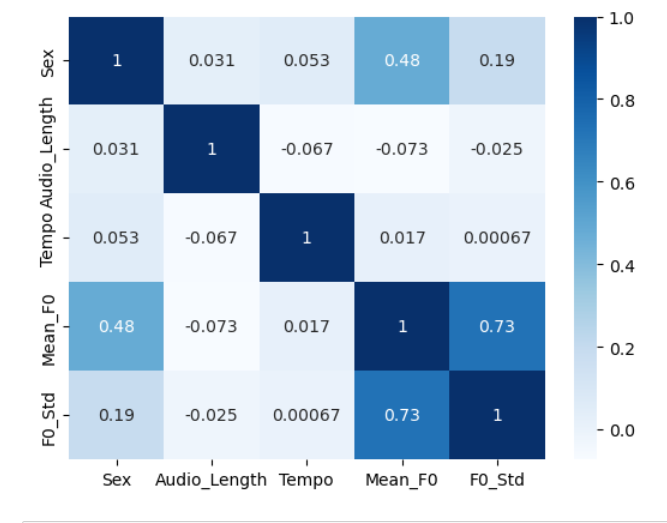


Figure 2. The confusion matrix of correlations between features and labels

As illustrated in Figure 1, there are two extreme outliers in the male group with mean F0 values higher than 1000 and even 4000. These outliers were normalized to the mean value of 153. Similarly, in the female group, there are two outliers with mean F0 values higher than 1000, which were normalized to 234. The descriptive statistics after normalization are shown in Table 1, with no changes observed in the statistical significance results.

**3. Gender Classification with Different Models**

I employed Logistic Regression, Random Forest, Multi-layer Perceptron (MLP), and Convolutional Neural Network (CNN) as four classifiers. The first three were trained with the extracted F0 features and are being compared, while MLP and CNN were trained with spectrograms and are also compared. Additionally, this report compares the performance of MLPs trained with two kinds of features.

**3.1 Extracted Acoustic Features**

**3.1.1 Logistic Regression**

Before training the model, the feature 'audio length' was dropped, while 'tempo' was retained despite its lack of significant difference between the two gender groups. Acoustic features extracted from audio recordings served as input to the classifier. The dataset was split into training and test sets, and the logistic regression model was trained using the 'multinomial' multi-class setting with 'lbfgs' solver, incorporating balanced class weights and a maximum of 1500 iterations.

The logistic regression model achieved a training accuracy of 82.32% and a test accuracy of 82.09% in distinguishing between male and female speakers. While the model exhibited slightly higher precision, recall, and F1-score for female speakers compared to male speakers, it showed balanced performance across both genders. The confusion matrix revealed 94 true positive instances and 137 true negative instances, alongside 32 false positive instances and 5 false negative instances.

**3.1.2 Random Forest**

A Random Forest classifier was trained and evaluated using the extracted features dataset, each labelled with binary classes representing male (0) and female (1) speakers. The dataset was split into training and test sets to assess the model's generalization performance. The Random Forest classifier was configured with 110 estimators, a maximum of 3 features per split, and a random state of 0.

The Random Forest classifier achieved an accuracy of 87.31% in gender classification tasks. Precision, recall, and F1-score metrics were computed to evaluate the model's performance. The precision of 84.62% indicates the proportion of correctly classified instances among those predicted as positive. The recall of 92.96% represents the proportion of correctly classified instances among all actual positive instances. The F1-score of 88.59% harmonizes precision and recall, providing a balanced measure of the model's performance. The confusion matrix analysis revealed 102 true positive instances, 132 true negative instances, 24 false positive instances, and 10 false negative instances.

**3.1.3 MLP**

The MLPClassifier achieved an accuracy of approximately 86.57% on the test set, indicating the proportion of correctly classified instances out of all instances in the test set. The loss computed by the model was approximately 0.3245, reflecting the model's alignment with the true labels during training. Confusion matrix analysis revealed balanced performance across genders, with 94 correctly predicted instances for male speakers, and 138 for female speakers, 32 male audios are predicted as female audios, and 4 female audios are predicted as male audios. While the model's precision for male speakers is commendable at 96%, its recall of 75% suggests some instances of male speakers being misclassified. Conversely, the model's precision and recall for female speakers are 81% and 97%, respectively, indicating robust performance in correctly identifying female speakers. The balanced F1-score of 0.84 for male speakers and 0.88 for female speakers further affirm the model's effectiveness across gender categories.

**3.1.4 Comparison**

Overall, Random Forest demonstrated the best performance across all metrics, followed closely by MLP. Logistic Regression, while still performing reasonably well, lagged slightly behind the other two classifiers in terms of accuracy and precision. Random Forest's superior performance can be attributed to its ability to handle non-linear relationships and feature interactions effectively, making it well-suited for this classification task.

However, these three classifiers' accuracy performances were quite close, with none achieving an accuracy higher than 90%. This was anticipated given the limited feature extraction. Additionally, all classifiers showed better accuracy in predicting female voices.

**3.2 Mel-Spectrogram**

**3.2.1 MLP**

The MLP classifier was trained and evaluated on spectrogram data for gender classification. The classifier achieved an accuracy of 76.12%, indicating the proportion of correctly classified instances out of all instances in the test set. However, the loss computed by the classifier was relatively high, measured at 0.7106, suggesting suboptimal alignment with the true labels during training. Further analysis revealed variability in the classifier's performance between the two gender classes. Precision for the male class was perfect at 100%, but recall was notably lower at 51%, resulting in an F1-score of 68%. Conversely, for the female class, precision was 68%, while recall was perfect at 100%, leading to a higher F1-score of 81%. This discrepancy in performance between the two classes indicates potential bias or class imbalance issues within the dataset. Overall, while the MLP classifier demonstrates moderate accuracy, further optimization and consideration of class imbalances are warranted to enhance its performance in this task.

**3.2.2 CNN**

The Convolutional Neural Network (CNN) architecture designed for the spectrogram-based classification task comprises several key components. Beginning with the input layer, the CNN processes spectrogram images, with each layer contributing to feature extraction and classification. Convolutional layers apply filters to the input, extracting spatial features such as edges and textures. Max-pooling layers downsample the feature maps, reducing computational complexity while preserving important features. The flatten layer converts 2D feature maps into a 1D vector, preparing the data for fully connected layers. These dense layers further process the extracted features, with ReLU activation functions introducing non-linearity. The output layer utilizes a sigmoid activation function to produce classification probabilities. Overall, this CNN architecture effectively processes spectrogram images, capturing spatial dependencies and hierarchical features to facilitate accurate classification.

The model achieved a remarkable performance on the test dataset, with an impressive accuracy of 97.01%. The model's test loss was measured at 0.0956, indicating minimal errors during the classification process. Precision, recall, and F1-score metrics were computed for both female and male classes. For the female class, the precision, recall, and F1-score were 96%, 98%, and 97%, respectively, while for the male class, these metrics were 98%, 96%, and 97%, respectively. Such high precision, recall, and F1-score values demonstrate the model's ability to accurately classify both male and female speakers. The macro-average and weighted-average F1-scores were both 97%, further affirming the model's balanced performance across classes. Overall, these results showcase the effectiveness of the CNN model in accurately discerning between male and female speakers based on spectrogram data.

**3.2.3 Comparison**

In comparison, CNN achieved superior performance with 97.01% accuracy, showcasing its capability to capture spatial dependencies and hierarchical features in spectrogram images for accurate classification. Overall, the CNN emerges as the preferred model for gender classification tasks based on spectrogram data, offering higher accuracy and robustness.

**4. Conclusion**

In summary, CNN with spectrogram-based training data showed the best performance among all the classifiers. MLP performed worse with spectrogram-based data than with extracted acoustic features. All three classifiers with acoustic feature-based training showed similar performances with better recognition accuracy for female voices, while the random forest classifier showed the best performance among them.