Classifying Gender Based on Audio Samples Using Machine Learning

CAB420 Assignment 2 Report

Sama Alkamachy - n10183787
Kevin Nguyen - n10269711,
Melissa Paton - n6334539,
Rafael Alves - n9683836

October 4, 2023

Executive summary

The following report will classify the gender of a speaker from an audio file. The model was trained on the LibriSpeech data set which contained audio samples taken from audio books along with relevant details such as speakerID and gender. A batch size of 10,000 samples of 5 seconds in duration was taken out of this data set and split into a training, validation and testing each containing an equal number of male and female speakers. The models used were, VGG (DCNN) with raw data and ResNet (DCNN), Random Forest, LSTM all using different feature extraction methods. The final results showed that the random forest performed poorly, whereas the DCNNs and LSTM were fairly similar reaching a performance of approximately 94%.

Contents

1 Introduction

This report will discuss the use of Machine Learning to classify the gender of the speaker in audio files. In order to create a model, it will be trained on the LibriSpeech data, which is a collection of samples taken from audio books. The research question being:

Can a Machine Learning model classify the gender of a speaker from only audio?

1.1 Motivation

Voice recognition is an essential aspect of modern human-computer interaction, enabling hands-free control and personalized experiences. However, voice recognition systems often lack the ability to identify the gender of the speaker, which can limit their effectiveness in certain contexts. By building a gender classification model from audio samples, we can enhance voice recognition systems by incorporating gender-specific adaptations. For instance, in virtual assistants like Siri or Alexa, accurately identifying the gender of the user can enable the system to respond with appropriate language and tone, providing a more tailored and natural interaction. Similarly, in call centers, gender classification can help route calls to the most suitable agents based on customer preferences, contributing to improved customer satisfaction.

While speaker identification or classification can also be relevant, gender classification specifically focuses on capturing gender-related characteristics from audio samples. Gender classification provides insights into gender representation and dynamics in various domains, such as sociolinguistics research or gender analysis in online discussions. Moreover, gender classification has practical applications in fields like marketing and advertising. For instance, companies can utilize gender classification to customize advertisements, products, or services targeting specific gender demographics. In the entertainment industry, gender classification can assist in voice casting for animated characters or generating voiceovers for media content based on the intended gender portrayal [1].

Overall, developing a gender classification model from audio samples serves as a crucial step in improving voice recognition systems, enabling personalized interactions, enhancing customer experiences, and facilitating gender-specific adaptations in various real-world applications, including virtual assistants, call centers, sociolinguistics research, marketing, and entertainment.

Additionally, by classifying the data rather than performing identification the models can be trained with a smaller data size as less samples are needed.

2 Related Work

2.1 LSTM

In this paper, the authors proposed to combine the Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) in parallel as lower networks in order to exploit sequential correlation and local spectro-temporal information [2]. The paper tested LSTM, CNN, Deep Neural Network (DNN) and mixed LSTM-CNN models on classifying audio collected from public areas (bus, beach, park etc). The audio was transformed into Mel Frequency Cepstral Coefficients (MFCC) features as input data for the model to extract the sequential information. Overall the paper found that the LSTM, CNN and Mixed LSTM-CNN models performed with better accuracy than the baseline DNN model. The order from worst to best was DNN, LSTM, CNN, CNN-LSTM models. Hence, LSTM and CNN were improvements to the paper's audio classification.

Researchers from the Chinese Academy of Sciences used LSTM for solar radio spectrum classification [3]. The researchers collected solar flare signals from the e Solar Broadband Radio Spectrometer (SBRS) in China, and predicted signals into 3 classes of solar "burst", "non-burst" and "calibration". The signal data was transformed into an image spectrum before model training. Deep Beliefs Network, CNN, Principal Component Analysis(PCA) + Support Vector Machine(SVM) and LSTM models were trained and tested on the solar signal data. The research found that the LSTM model performed the best with the highest True Positive Rate and the Lowest False Positive Rate. The worst model performed was the PCA+SVM model.

2.2 Random Forest

Authors Zhang and LV used Random Forest to train and test audio classification on Environmental Audio data, classifying classes like Birds, Wind, Frog etc [4]. The audio data was set to a sample rate of 8kHz, a length of 10 minutes and all silent noise was removed. The audio was then tested for types of data feature representations like MFCC and Code Excited Linear Prediction (CELP). Models Bagging and AdaBoost were also tested and compared with the Random Forest. The paper found that Random Forest significantly outperformed Bagging and AdaBoost Models in most cases, and that MFCC could obtain better prediction performances than CELP. This paper shows that Random Forest and MFCC had good performance.

Researchers at Oxford University used Random Forest and multi-way SVM classification algorithms for image classification uses [5]. They classified images with the object categories (eg. dog, flower etc.). It was demonstrated that the Random Forest and SVM algorithms performed with similar results (at 80% and 81% respectively), however, Random Forest reduced training and testing costs significantly over multi-way SVM. This shows that Random Forest performs well when wanting to reduce training time and suggests that Random Forest is good at image classification. This is important as our model will essentially be using image classification of audio images (MFCC, log-Mel Spectrograms) to classify the data.

In their paper, Vimal et al. detail the process of extracting MFCC and audio energy as the main parameters which are used by Decision Tree, Random Forest and SVM models for emotional classification of audio[6]. Using the aforementioned features it was found that the random forest algorithm produced the highest accuracy classification with an accuracy of 88.54%.

2.3 VGGnet

The study uses retinal images from diabetic patients for image classification to assess whether a patient has retinal disease [7]. Each image was classified based on the severity of the retinal disease, ranked from 0, 1, 2, 3 and 4 (with 4 being most severe). The models tested were Alexnet, VGGnet-s, VGGnet-16, VGGnet-19, GoogleNet and ResNet. The results of the study were that GoogleNet performed the best when the parameters were randomly initialised, VGGnet and Alexnet performed similarly with GoogleNet and ResNet performed very poorly with 0.7% accuracy. However, when the authors applied hyperparameter-tuning for the parameters, VGGnet-19 performed the best with other models being similar, and the worst model AlexNet getting 89% accuracy for hyperparameter-tuning. This study shows the reasonable efficacy of VGGnet and ResNet at classification tasks.

The researchers tested CNN models with audio classification of music genres. The data for the models were represented The models tested were VGGnet, 1 dimensional waveform CNN, MFCC+SVM etc. The researchers found that VGG performed the best as it had the highest accuracy out of all the models, suggesting that VGG performed well as it was good at identifying spectral features. VGG was also good at identifying timbre (tone) of the music audio.

2.4 ResNet

Authors used DCNN, SVM, RF and k-Nearest Neighbors(KNN) models to train audio classification sounds of patient coughing to detect Covid-19 [8]. A Mel-like spectrogram was applied on the data to capture time frequency information. The models had minor alterations applied such as AUCO ResNet, Shallow Learning. The study found that AUCO ResNet performed the best with the highest accuracy. Shallow SVM, RF and KNN performed with significantly worse accuracy. Hence, the paper shows that the altered ResNet model performs significantly well against the altered SVM, RF and KNN models at identifying Covid-19 using coughing audio.

2.5 Features

This paper explores the effectiveness of log-Mel Spectrogram (shown in Figure 1) and MFCC features for Alzheimer's Dementia [9]. The paper uses data from the ADReSS dataset that consists of speech recordings of participants, balanced for age and gender to minimise risk of bias in the prediction. The models trained were CNN-LSTM, Resnet-LSTM and pBLSTM-CNN, and the input data was represented as either log-Mel Spectrogram or MFCC. The study found that CNN-LSTM MFCC performed better than the Mel-Spectrogram as it had a better accuracy and lower RMSE score. The Resnet-LSTM was only trained for Mel-Spectrogram and it performed very similarly with CNN-LSTM and better than pBLSTM-CNN. The pBLSTM-CNN MFCC had a large accuracy performance increase and smaller RMSE value from the Mel-Spectrogram. This means that for CNN-LSTM and pBLSTM-CNN, MFCC representations of the data performed better than log-Mel Spectrogram representations at classifying the speech data from Alzheimer patients. This may suggest that MFCC is a feature that minorly performs better for audio than log-Mel Spectrogram.

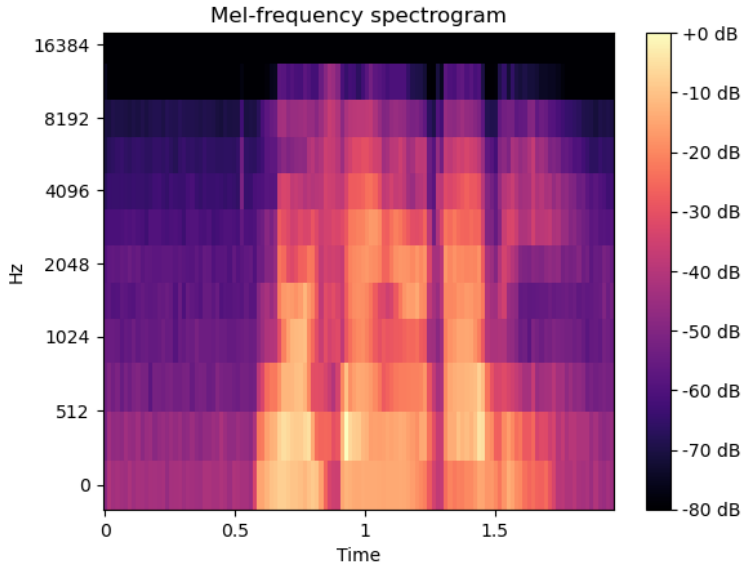

Figure 1: Visual representation of a log scaled Mel Spectrogram

2.6 Data Augmentation

In paper 2, the authors explored a new method of data augmentation for audio classification called Mixed frequency Masking data augmentation that could reduce overfitting and improve on the generalisation ability of deep neural network [10]. Pre-processing changed the audio data into a log-mel spectrogram as in previous literature it suggests that log-mel spectrograms had the best performance as the input of convolutional neural networks than other audio to image techniques. Then a variety of data augmentations were applied to the log-mel spectrogram dataset such as: add noise, time stretch, mixed frequency masking etc. The augmented data was then inputted and trained in a Resnet model and tested on a testing set. Each data augmentation method was tested 5 times and the accuracy was averaged. Overall, the authors new data augmentation Mixed frequency Masking successfully reduced overfitting and improved accuracy on the baseline accuracy of the Resnet audio classification by 1.28% accuracy.

2.7 Dimension Reduction

The study used raw 1D and MFCC audio data of normal and abnormal sounding hearts, collected from patients using a stethoscope, to then train and test on gradient boosting algorithm, random forest and SVM models [11]. The study applied noise filtration on the dataset to remove background hospital noise and increase quality of the heart sound. Then the study applied dimension reduction on the audio data as it was shown in previous studies that Linear Discriminant Analysis(LDA), PCA and Genetic Algorithm(GA) improved accuracy of the audio classification model. This is because the performance of the model degrades when too many dimensions are included into the input, hence dimension reduction reduces the degradation of model performance and improves accuracy by only selecting the important features. The results of the study found representing the data using MFCC gave approximately 20% increase in accuracy from 1D raw representations of the audio data. The study also found that using dimension reductions PCA, LDA and GA did not add to accuracy of the 3 models but decreased them.

3 Data

The audio files used for this analysis were from the LibriVox project, accessed through the Librispeech library (specifically the train-clean-100 dataset). This dataset contains 1000 hours of audio, sampled at 16kHz, comprising a total of 28539 samples (each about 15 seconds in size). The files were samples from public domain audio books. There were 251 speakers (126 males, 125 females) in the dataset, each audio file contained the speakerID in the name, additionally a "SPEAKERS.TXT" file was provided that had the gender of each speaker. The audio files were clear, intelligible and with minimal ambient noise. The dataset also contained the transcript; however, this was not used for the analysis.

The audio files were loaded using librosa, a python audio library, transforming the data into an audio time series, which is a one-dimensional array capturing the amplitude at the different time points specified by the sampling rate as shown in Figure 2. The dataset size was reduced to a batch size of 10,000 due to the limitations discussed in Section 3.2. To ensure that the dataset is balanced across both male and female genders, the dataset was split evenly across the number of speakers per gender and the number of samples per speaker. Therefore, the dataset used was 10,000 audio files with an equal amount of 125 male and female speakers that have 40 samples each.

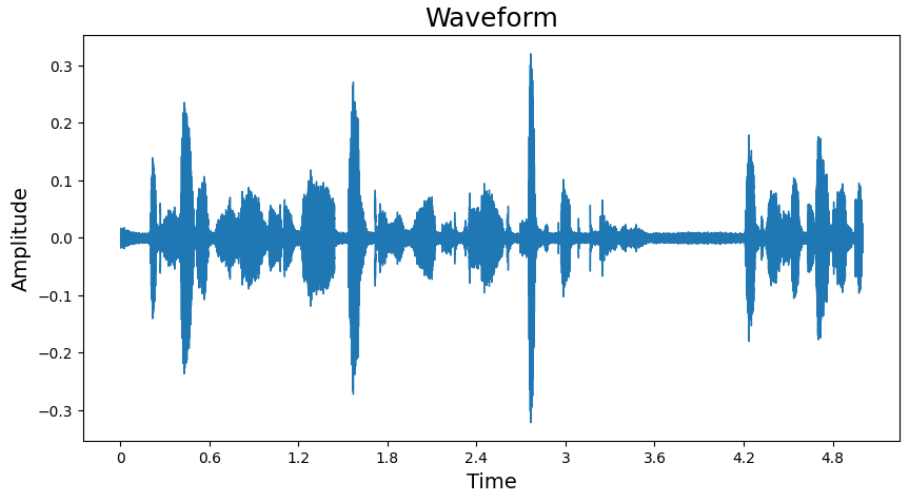

Figure 2: Audio waveform of one of the audio samples

Additionally, the LibriSpeech data is a zipped file which contains multiple folders inside of it. The main audio folder contains sub-folders which are the speaker ID, within those folders is the .flac data needed. Loading the audio files into the code, the sequencing of the audio files has a pattern, i.e. ['M','M','M','F','F','F'...]. The data was shuffled to avoid the models from training any sequence patterns. The shuffle random seed was set to 42 to maintain consistency throughout training and testing across all models.

The data set was split into a training, validation and testing set from the main data set using a 40/30/30 split. Each set had an even distribution of male and female samples as seen in Figure 3.

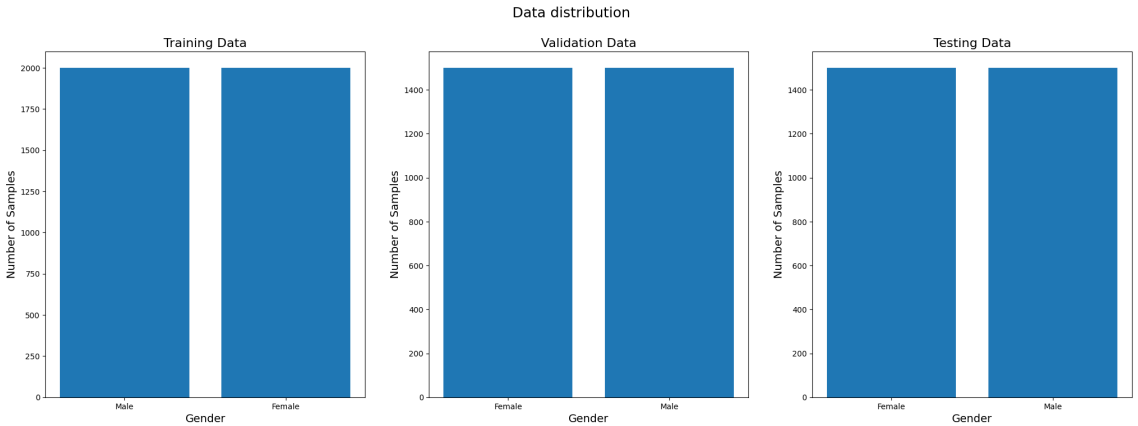

Figure 3: Data distribution of the batch taken out of LibriSpeech.

3.1 Generic Pre-processing

The following pre-processing was performed for all models. The generic pre-processing for the audio classification included:

1. Each audio file in the model is a different length, making it a dynamic input size. There are many methods to handle dynamic input size, some commonly used methods are padding, chopping, sliding window. Due to computational limitation the method chosen was cropping. Therefore, a standard sample size of 5 seconds was chosen, any samples greater than 5 seconds was clipped at 5 seconds those under 5 seconds were removed (as such, padding was not required).

2. When capturing audio, it was important to sample the data at a rate that was higher than twice the highest frequency present in the audio. This is known as the Nyquist-Shannon sampling theorem. Sampling at this higher rate helps prevent a problem called aliasing. Aliasing happens when high-frequency parts of the audio get mixed up with lower frequencies, causing distortion and loss of information. By sampling at a rate that is at least twice the highest frequency, we can accurately capture and reconstruct the original audio without aliasing [12]. The data is originally 16kHz, it was loaded into the system using a sample rate of 32k.

3. The gender was transformed to a numerical representation for binary classification. Audio samples that were a male speaker were represented with 1 and female speakers with 0.

The final outputting audio samples (x data) is 10,000x [160000,1] audio samples split into training, validation and testing using a 40/30/30 split. With the y data (labels) representing the gender as 0 and 1 in a [10000,] array.

3.2 Limitations

The main limitation the project faced was the lack of computational resources. The machine learning models often demand high computational power for processing, loading data, and training models. The data loading had to be limited to 5 second clips of the original audio samples which could range into 15 seconds or more. While the data was of uniform frequency bandwidth, if the data had higher frequencies it might have been required to filter out the higher signals rather than use a larger sample rate.

The audio data set has a good balance of male and female speakers, but individually the speakers can have varying numbers of available data by having shorter audio files or just less files in general. Care was taken to have a balanced dataset, but this meant some speakers could have their datasets cut in half in order to prevent contributing to class imbalance.

In addition, all the samples in the dataset are categorized strictly into two genders: male and female. Consequently, the model's classification capabilities are restricted solely to these two genders. It's worth noting that these samples exclusively consist of audio recordings from narrators of audiobooks, and non-English accents are not included. Since various languages influence the human voice in distinct ways, this limited dataset realistically confines the model's predictions to native English speakers.

The dataset offers the model a valuable advantage of clean audio data, enabling it to extract features with exceptional accuracy. However, it is essential to consider that real-world applications often involve audio data captured from devices like Alexa, which may contain significant amounts of noise or involve multiple speakers. As a result, the absence of noise in the training data further restricts the models' capability to effectively handle and separate noise in practical scenarios.

4 Methodology

There were 4 methods chosen to categorise the gender of the speaker given the LibriSpeech data set.

4.1 VGG network with raw data (DCNN)

There has been no research on using raw data samples in a VGG network to classify the gender of the speaker, but there has been alot of research into using VGG networks for other classification tasks using raw data where it has been proven to be efficient [13]. However, gender classification tasks are complex, as there is alot of diversity in the human vocal range; with humans themselves failing to categorise the difference between male and female audio at times.

A VGG model was used on the raw audio data. The raw data is a 1 dimensional representation of the amplitude over time, therefore rather than using a 2D convolution layer a 1D convolution layer was used. Other than that, all the layers remained as they are with any other basic VGG network: A convolution layer, followed by a max pooling layer, which down-samples by pruning the input, repeated 3 times over 16, 32 and 64 filters then flattened and passed through a dense layer. Up until the dense layer the network is extracting features from the input and compressing it - Essentially acting as a feature extraction method. The final dense layer is then passed through a sigmoid layer which squash the predictions between 0 and 1. The Gender labels are already numbers, with male being 1 and female being 0 so a binary cross entropy loss function was used.

The full model architecture is shown in the Appendix (Figure 7). The model found over 5 million trainable parameters.

A call back was added to the training of the model, the call back stops the model when it has converged. The model has converged when the accuracy or loss has stopped improving. The call back has a patience of 5 attached to the validation loss, that means when the validation loss has not improved in 5 sequential epochs then it will stop the model from training. This is used to ensure the model has learned all that it can from the data without having to pick an arbitrary epoch number to test and inspect.
The model was trained on 100 epochs, and stopped at epoch 21.

A batch size of 128 was chosen for training the model, this was chosen through trial and error.

As discussed in [13], DCNN models require a variety of data to accurately extract features. Furthermore, the task (classifying gender) is complex and would require a wide range of diverse audio samples to accurately work in real life application. Therefore, data augmentation was introduced to the raw data to improve the quality of the data as this has been proven to be successful [10]. 1% of the training data was randomly chosen and augmented slightly. The augmentation was kept simple as to not corrupt the data; the pitch was slightly shifted, and random noise was added. Each augmented sample is transformed randomly, ensure the augmentation is not a constant. While augmentation can be very helpful for models it can also be damaging if too much is added. The audio for samples with noise were validated to confirm that noise didn't detract from the speaker.

4.2 Resnet with Feature Extraction (DCNN)

The Resnet model was chosen to be an appropriate model for audio classification tasks as it was used and recommended for audio classifications by Wei et al, 2019 [14]. The audio of the data was transformed from a 1-dimensional signal into a log Mel Spectrogram figure representation of the audio. This would create a 2-dimensional representation of the data that performs better than raw audio data representations [15].

The Mel Spectrogram is a 2-dimensional plot representation of the audio data, that can capture both high and low frequencies at the same time. The amplitude of the sound is also scaled by the mel scale which mimics the sound that a human hears by making low frequency sounds larger in the plot and high frequency sounds smaller in the plot. This is because human ears can distinguish the difference between 500Hz and 1000Hz easily but struggle to distinguish 7500Hz and 8000Hz [15].

The Mel Spectrogram underwent further pre-processing by being resized to 128, 128 pixels and gray scaled to 1 colour dimension. The image was resized due to computational limitations and the image was gray scaled as the image already followed a 1 dimensional colour axis, but gray scaling removed the rgb colours. The Resnet model architecture was a Resnetv2 model as Resnetv2 performed better than Resnetv1 in the DCNN Example 5 Resnet notebook. There are 3 residual blocks in the model consisting of 2 convolutional layers. Each

convolutional layer had a batch normalization layer to allow the nodes to be values between 0 and 1 letting the model train faster and more stable, followed by ReLu activation layer to remove all negative inputs. The Resnet model uses a skip propagation that addresses vanishing gradient problem, where the value of gradients diminish as it travels through layers during back propagation. Skip propagation allows the model to learn faster by allowing a shorter path for data through the model.

The Resnet model used an input of 128, 128, 1 as the image sizes were 128, 128 with 1 gray scaled colour. The model uses residual blocks with 16, 32 and 64 filters. These filters were chosen to increase parameters in the model and help the model learn key information from the image input information. The output layer was 1 dense layer with a sigmoid activation layer that forces the model to exclusively predict female (0) or male (1). The training of the model uses binary cross entropy loss function as the model has a binary output. The batch size of 64 was chosen as batch size 128 and batch size 16 both encountered overfitting too quickly at epoch 5 with low accuracy, whereas batch 64 outperformed both batch sizes. The model trained for 10 epochs as the mode would overfit and the training accuracy would hit 100%, hence the model did not need to train any further.

4.3 Random Forest

A Random Forest model was chosen to be used for speaker gender categorization to take advantage of the many possible parameters that can be generated from audio data. After the dataset had been created from the audio files, it could be further processed into several parameters which are then collated in a dataframe. The parameters are:

- **Mel Spectrogram**
- **MFCC**: A Mel-Frequency Cepstra Coefficient is a Mel Spectrogram that has had an Inverse Discrete Fourier Transform applied to it which the result of is seen in 4 compared to its original form in 1. This returns several coefficients which are essentially the several components of the Mel Spectrogram, these coefficients are related to specific components of speech or sounds. The first dozen components are commonly used for audio classification tasks related to speech and as such are what the model uses for its MFCC parameters rather than using all MFCC components.

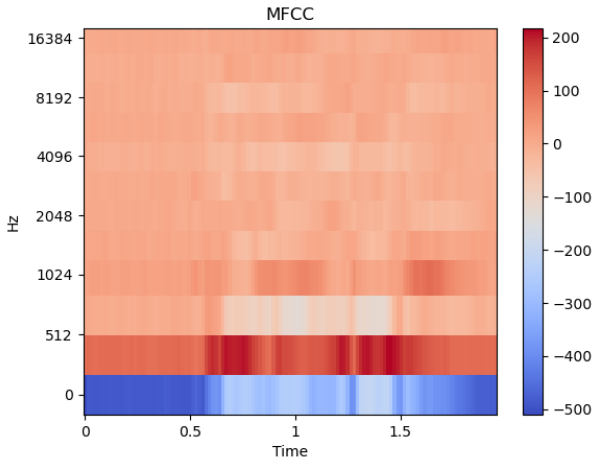

Figure 4: Mel-frequency cepstral coefficients

- **PCEN**: Per-Channel Energy Normalisation is a normalized time-frequency representation of the audio signal that has been transformed by gain control and nonlinear compression. This has the effect of reducing the dynamic range of the audio which helps reduce noise and focus on the desired components of the audio, in this case: human speech.
- **MFCC Delta**: An important factor in audio analysis of human speech is the rate of change in the amplitude of the signal. MFCC delta is a method to quickly approximate the derivative of the signal by subtracting f(t2) by f(t1). While quick and computationally light, this method is susceptible to noise in the signal, the delta of the MFCC delta can also be calculate and is doubly vulnerable to the noise.

The parameters are loaded into a dataframe in order to be used by the random forest model. The Random Forest Classifier itself has several parameters that should be adjusted to both optimize the time to create the

model and improve its accuracy. To choose the optimal hyperparameters a Grid Search cross validation method was used with k-Fold value of 3. Using this method the following hyperparameter values were chosen:

- **Max Depth**:20. This is the max depth the tree will descend to, the default behaviour is to continue until the leaf nodes are pure or leaf nodes contain less than the minimum samples per leaf parameter.
- **Max Features**:3. This is the number of features the algorithm will consider when trying to create a split, if needed it will go above this setting if that's required to create a valid split.
- **Max Samples**: 0.6. This is the amount of samples to be drawn from the training data to train an estimator. At 1.0 that means the entire training data will be considered per tree, by lowering it we introduce some randomness into the system which may result in training on different features and reduce overfitting
- **Min Samples Leaf**: 1. The minimum samples required for a leaf node to exist. With a value of 1 this has no change from default tree modelling.
- **Min Samples Split**: 4. The number of samples required to split a node, if there's less than 4 samples available the tree will not be split.
- **Estimators**:1500. The amount of estimator trees that will make up the random forest model.

The random forest model was set to use a random state, which would control random variations in its training when retraining the model. The Gini Impurity method was chosen for the split criterion, the Gini method is computationally lighter than the Entropy method so even though it has a slight loss in accuracy it was a negligible loss for improved train times in development.
Out-of-bag Error was used in the Random Forest model. OOB error is essentially running cross validation on the remaining third of the data which isn't being used in the decision tree to estimate the generalization score. Unlike cross validation, OOB is unaffected by data leakage, but as the dataset and parameters grow it can become resource intensive to compute. As the number of estimators is increased the OOB scores begin to lower and plateau, providing a good range of estimators to minimise the generalization score.

4.4 Long Short Term Memory (RNN)

Long Short Term Memory model was developed to categorize the gender of the speaker. This model was selected as it has been proven to be effective at classifying speaker's gender [16]. Audio data is time series data and in the case of this model, each sample contained 160,000 data points per sample. The recurrent nature of an LSTM means that past patterns can be extracted from the feature space, instead of analysing each data point for a sample in isolation. Depending on the window size, these patters may be detected over long periods of time.

Mel Spectograms were used as inputs to the LSTM models (see the Resnet model for justification). The Mel Spectograms had a hop length of 512, and the window length for calculating the Fast Fourier Transformation was 1024. 64 Mel Spectogram bins were used. The final shape of the input (after being transposed) was (313, 64).

Architectures investigated included 1 layer, two layer, three layer and stacked LSTMs. For each model, a sigmoid activation was used for the final dense layer (to return a value between 0 and 1). Over stopping and model checkpoints were trialed to reduce training time, however these significantly reduced performance. Overfitting wasn't observed for any of the models, so early stopping was also not required. Adding an attention layer was also trialed, however this did not improve performance. A batch size of 100 was used due to performance and computational reasons. Different batch sizes were trialed however they took too long to train or decreased model performance. LSTMs with additional layers and deeper stacked LSTMs were also trialed however this did not increase performance. The loss from the stacked LSTMs were also not as stable as the other LSTMs.

The model with the best architecture was the single layered LSTM (considerin performance and training time). This model has with a hidden dimension of 70 and 16 filters. This model was trained for 100 epochs, with the accuracy and performance above 0.94. This model trained in 1575 seconds, with an inference time of 3.5 seconds.

As there was minimal improvement observed when adding a second layer, and the performance declined significantly when adding a third layer, there may be an issue with a vanishing gradient. The stacked LSTM also has a lower f1 score and accuracy that the single layer LSTM. LSTMs are intended to address the vanishing gradient issue by accessing the forget gates of the LSTM via an additive gradient structure, with frequent gate updates with each model step the network can ensure that the error gradients do not approach zero. The models were continuing to learn, however at a slow pace, which supports a vanishing gradient issue. Further investigation would be required to support this theory and implement a fix.

5 Evaluation

5.1 Model Training - Neural Network Methods

Shown in Figure 5 is the epoch performance on the validation set for the 3 neural networks. The training performance was taken out of the plot as to not crowd the figure. The Random Forest does not have any epochs as it does not use gradient descent to train, instead it uses several methods like adjusting for out-of-bag error or hyper-parameter tuning.

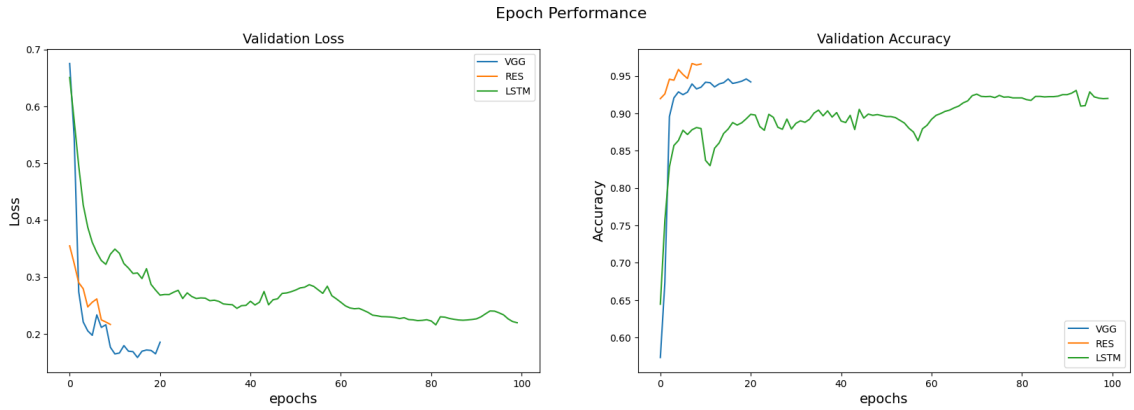

Figure 5: Validation set Epoch performance across the 2 neural networks.

This model shows the different behaviour of the neural networks. The ResNet is starting its first epoch with an accuracy greater than 90% whereas the other models are slowly building up to it than converging. From this graph, the ResNet has not actually converged yet and requires more epochs to converge. The ResNet model is also using a spectogram which means the features are already extracted from the audio before passing it to the model; which is why it starts at a high accuracy. The VGG model is using the raw data which is why the accuracy starts low and builds up - it requires time to extract and learn the relevant features. Similarly the LSTM while also using feature extraction still requires time to find patterns in the data recursively.

5.2 Model Training - Random Forest

The hyper-parameters of the model were fine tuned using a grid search approach, with a k-fold value of 3. For the first grid search, the parameters were chosen by selecting a minimum and maximum value based on the data inputs to the random forest model eg. the max depth cannot be arbitrary as it is reliant on the length of the data, max features will be capped at the amount of features in the training data. Using the output of that initial grid search the range of the values were lowered to be centered around the best parameters in order to more finely calculate the best parameters. In the end the parameters for the Grid Search and their ranges to test were: 'max_depth': [10, 20, 30, 40, 50], 'max_samples': [0.3, 0.6, 1.0], 'max_features': [1, 2, 3, 4], 'min_samples_leaf': [1, 2, 3, 4], 'min_samples_split': [2, 4, 6], 'n_estimators': [200, 300, 500, 1000, 1500]

5.3 Performance

All four models were ran on the same computer using the same data set for a fair evaluation. The performance of the models is captured in table 1 below. All of the models are outputting a number between 0-1, therefor a threshold must be added to classify the predictions as one or the other. A threshold of 0.5 was chosen, this means if the model outputs a prediction of 0.6 on a sample that is classified as 1 (male).

Model Performance				
Model	**Training Time (sec)**	**Testing Time (sec)**	**Accuracy (%)**	**F1 Score**
Resnet	1809.2	264.7	96.6	96.6
VGGnet	2476.3	164.7	94.4	94.6
LSTM	1797.7	7.6	94.3	94.4
Random Forest	202.3	0.63	58.0	58.2

Table 1: Model Performance Table

The model training time shows that the Random Forest performs the best out of all the models with the shortest time (202.3 seconds), whereas the other 3 networks had significantly larger training times, approximately 10

times larger than Random Forest. Faster training times is typical of non-neural network models.

The testing times show that Random Forest reduces the testing cost of the model significantly to the other models at 0.63 seconds. The LSTM also outperforms its DCNN counterparts significantly. The DCNN models VGGnet and ResNet have large testing times at 164.7 and 264.7 seconds respectively.
The VGG net has the longest training time - 2x the ResNet, due to it using raw data. The VGG net found over 5 million parameters whereas the ResNet (using feature extraction) found 130k.

The DCNN models have very similar Accuracy scores with ResNet performing the best with 96.6% accuracy and VGGnet and LSTM performing 94.4% and 94.3% accuracy. The Random Forest performed significantly worse than the other models. This means that while the Random Forest trains and tests significantly faster, the Random Forest did not efficiently and effectively learn from the training set data compared to the other networks.
It should be noted that the Random Forest method was originally chosen when the scope was speaker recognition as that would have had many categories, by swapping to gender identification the task changes to binary categorization. A different classifier might have been more appropriate to use given the change in focus like an SVM classifier utilising dimensionality reduction.

No insight can be gained from the F1 score as the dataset used is balanced and the F1 scores are very similar to the accuracy.

5.4 Model Comparison

The confusion matrices for the models can be found in Appendix(8).

Confusion Matrix				
Model	**True Male**	**False Male**	**True Female**	**False Female**
Resnet	0.98	0.053	0.95	0.019
VGGnet	0.95	0.068	0.93	0.053
LSTM	0.89	0.035	0.97	0.11
Random Forest	0.61	0.47	0.53	0.39

Table 2: Model Confusion Matrices

As seen in Table 2, Random Forest performs worst in all categories with low True values and high False values meaning that Random Forest is not accurate. Both DCNN models, ResNet and VGGnet have similar results in each case, but with ResNet performing minorly better. The LSTM model performs worse at predicting True Males than the DCNN models, however the model performs better than the DCNN models with True Female. As recurrent nature of the LSTM and the Mel Spectogram may identify that the female voice has a consistently higher frequency than the male voice.

Both Resnet and LSTM models have mechanisms to handle vanishing gradients (skip propagation for the Resnet and forget gates for LSTMs). These two models use the Mel Spectograms in comparison with the VGG Network that uses the raw data. Ultimately as female voices have a higher frequency then male voices, the recurrent nature of the LSTM may not be required to detect gender in comparison to other neural networks.

6 Discussion

6.1 Failure cases

All models except the LSTM performed worse at classifying females than males - while this difference is very minor (0.03 between male and female), the fact that its replicated in 3 separate models could indicate a bias in the data.

The wrongly predicted data in each model was stored in an array, then the audio samples that were incorrectly predicted in all 4 models were inspected. There was a total of 6 samples that are incorrectly predicted in all 4 models. When listening to the samples the following characteristics were noted.

1. Loud echo
2. Dramatized voice
3. Nothing unusal noted.
4. a child.
5. Dramatized voice
6. Nothing unusal noted.

Through the above list, the first obvious failure in the models is predicted children. This is expected, children have not been through puberty, so the vocal chords have not changed yet. The second failure case noted are dramatized voices, these samples are from audio books which means most of them are more normal. These cases include male voices using a higher pitch (higher frequency) and acting more emotional. All 4 models are struggling with this aspect.

6.2 Comparison with external models

Facebook's Wav2Vec2 XLS-R model, a Convoluted Neural Network (CNN), was trained using five different datasets (Common Voice, Multilingual Librispeech, VoxLingua107, BABEL and VoxPopuli) to perform Speech recognition, translation and classification tasks. This model has been fine tuned to perform gender classification. The model had an f1 score of 0.9993 and loss of 0.0061 when classifying gender. The underlying model was trained on a broader training set, with a different train/test split, including 128 languages. The model also has 300 million parameters, making it a much more complex model than the models discussed today (LSTM has 39000 parameters, ResNet has 140,000 parameters and the VGGnet model has 5 million parameters). The Wav2Vec2 XLS-R model was trained on over 436K hours of audio. This model was trained on nearly 14 ours of samples (10000 samples, each 5 seconds long). As such it's not surprising that the model performance for this model greatly exceeded all of the other models presented.

On an emotion classification task based on audio clips, a random forest model performed the best out of all the classification models they used with an accuracy score of 88.54[6]. They used the RAVDESS dataset which consists of 24 actors with a 50/50 gender split, each with 60 audio files for a total of 1440 files (42.8GB). Although they were performing emotion classification which consisted of 8 target emotions, the random forest model still topped out at a sub-90% accuracy score where a hybrid network model using DCNN and LSTM achieved an accuracy of 98%[17].

This could be a sign that random forest itself might not be appropriate for this task while only using feature extraction, it might need to be combined with dimensionality reduction or be part of a hybrid model to become competitive with deep learning methods.

6.3 Improvements

All the models (except random forest) are already performing well, as shown in Table 2. However, due to the lack of computational power, models were developed, trained, and tuned on smaller batches of data so that development wouldn't be interrupted by 20 or more minutes of waiting for training. In the end all models were trained and tested on the best available computer with a much higher batch size, but the first time the models were used on the larger data set they produced significantly different output than ideally simply scaling in accuracy with more data available. This required team members to re-tune models without quick access to training and testing on larger dataset, given how much the team relied on repeated runs to test performance during development this could have impacted the fine tuning of the models to the data set that will actually be used.

To address this, data augmentation techniques can be employed to introduce noise and diversify the audio samples, potentially reducing the performance gap between male and female classification and improving the Random Forest's performance. Currently, only the VGG model uses data augmentation, and it was observed that without augmentation, its performance significantly decreased. The non-augmented data also exhibited a larger performance gap between male and female classification as shown in Figure 6. To improve the model performance of all four models, increasing the number of samples, the length of samples and increasing model complexity may improve model performance (as per the Facebook model [18]).

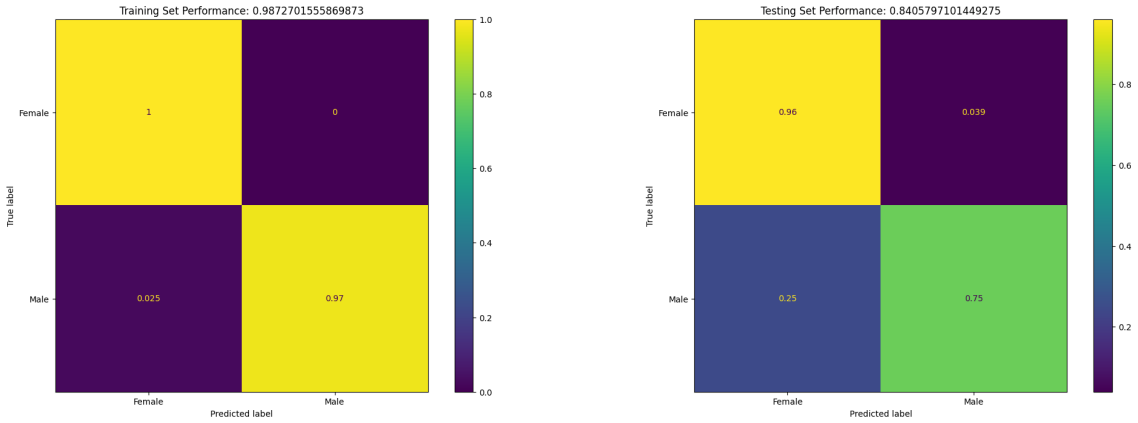

Figure 6: VGG confusion matrix, with NO data augmentation

Furthermore, combining a DCNN network (such as VGG or ResNet) with an LSTM model can enhance gender classification performance. This approach, which has shown promising results in related work, was not attempted due to time constraints, limited computational resources, and lack of expertise in the subject. Exploring this combination may lead to further improvements in the accuracy and robustness of the models.

MFCCs and Mel Spectrograms are regarded as the standard and one of the best parameters for audio tasks, especially speech related ones, there are many parameters that were not used which could have benefited models. Even with the MFCCs the coefficients were limited to 12 but the MFCC process can return more than just 12, the first dozen are usually regarded as most useful in speech analysis while the later ones give more information about non-speech related noises but one weakness of the models were samples with echos or voice distortion so maybe that could have impacted learning.

Energy parameters like Average Energy, Low Short Time Energy Ratio, and Short Time Energy could have been used; zero crossing rate provides potentially unique insight by measuring how many times the frequency crosses the 0 axis; spectral flatness, centroid, rolloff, and entropy are all common sound parameters which might have been useful given the MFCC already makes use of the spectral envelope. All of these merit further testing and were not implemented due to time constraints as careful research into their impacts would be required as they might emphasize non-human speech noises especially in distorted or noisy recordings if data augmentation or more datasets were used.

The creation of the dataframe could also be a point of improvement in regards to how the parameters are represented in a way to be processed by the model training. Some projects have been observed reducing a parameter that is thousands of data points long into a single mean or median parameter, while others simply concatenated the values. More time would be required to test the impacts and performance of different dataframes and potentially combining it with dimensionality reduction or hybrid models could improve performance.

7 Ethics

The dataset LibriSpeech used for the training of the model was collected from clips of speaker's audiobooks. There is no mention on the LibriSpeech website of the speakers providing consent of their voice to be used for machine learning research, hence there are ethical concerns for the privacy of the speaker and whether they are willing to participate in machine learning research.

The audio dataset collected was sampled from clips of English-speaking audiobooks. The types of speakers that contribute to the audio dataset could over-represent English-speakers from Anglo countries, under-representing the accents of minority ethnicities with English as their first language (Caribbean, Singaporean, Scottish etc.). This could lead the audio classification model to be less accurate at identifying the separation of male and female voices in these ethnic minority English-speakers, reducing the overall usefulness of the model for certain English-speaking demographics.

The models used separate male and female voices into two classes. Modifications could be made to the model to include further classes that separate types of voices along different identity accents. These modifications could be used to identify groups and maliciously discriminate against certain identity groups.

Furthermore, it is essential to recognize the ethical implications of classifying genders solely into male and female categories based on the available data. This binary classification may overlook the experiences of individuals who do not identify strictly with either gender or fall outside the traditional gender binary. By solely relying on such limited categories, there is a risk of reinforcing societal norms and excluding individuals who identify as non-binary, gender queer, or have other gender identities. A broader gender spectrum needs to be considered when capturing data to ensure that models or treatment regimes can be determined for all people [19].

Machine learning models are complex iterative models. This means that it is difficult for researchers to ascertain reasoning for how hidden layers of the model produce certain outputs. These hidden layers could create inherent biases within the models predictions that are difficult to justify, potentially leading to unjust model predictions with no clear solution to resolve it.

8 Conclusion

The model with the best performance was the Resnet model, with an f1 score of 0.966 and an accuracy of 0.966. Recommendations for future work include combining an LSTM and DCNN models to identify gender and additional data augmentation (including noise). Increasing the number of samples in the data set, including samples from people with a non-English speaking background and including samples from a broad age group is also recommended. These will address limitations identified in model performance.

Future considerations could consider using the age of speakers as an added feature to address the failure case presented in section 6.1. Classifying the genders into 'Female child', 'Female Adult', etc, has proven to perform at 99% accuracy [20]. Therefor, adding the age is likely to improve the performance and address the failure case.

9 Contributions

Kevin Nguyen - wrote code for ResNet, Worked on 10 Related Works papers, ResNet Methodology, Ethics, Model Comparison. Minor changes in Data. **25%**

Sama Alkamachy - wrote code for VGG, wrote code for loading in the data, ran all models and recorded performance, VGG methodology, Introduction, Data, Evaluation, Discussion . **25%**

Melissa Paton - wrote code for LSTM, LSTM methodology, Some of the Data, Evaluation, Discussion, Conclusion. **25%**

Rafael Alves - wrote code for Random Forest, Random Forest methodology, RF model training, RF external model comparison, old Related Work†, addition to improvements **25%**

10 Appendix

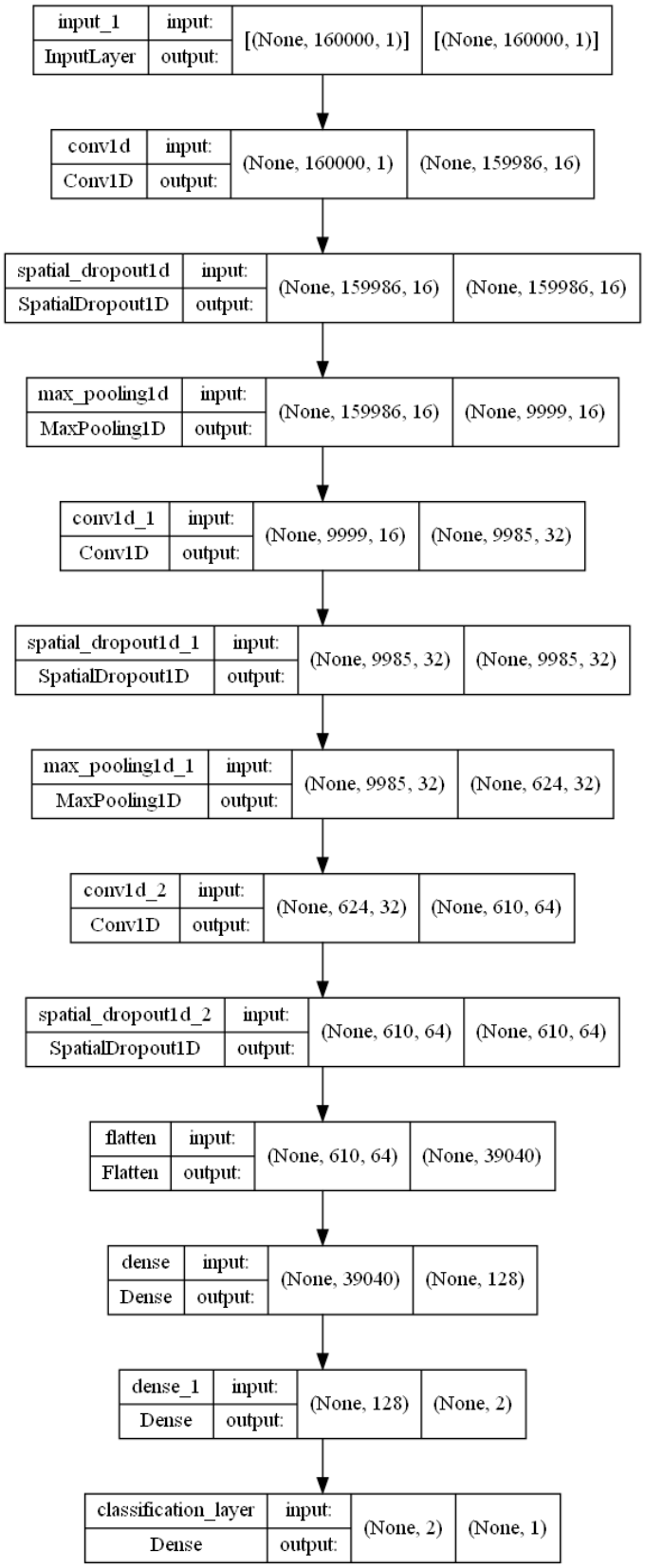

Figure 7: VGG network, with raw audio data input, model architecture

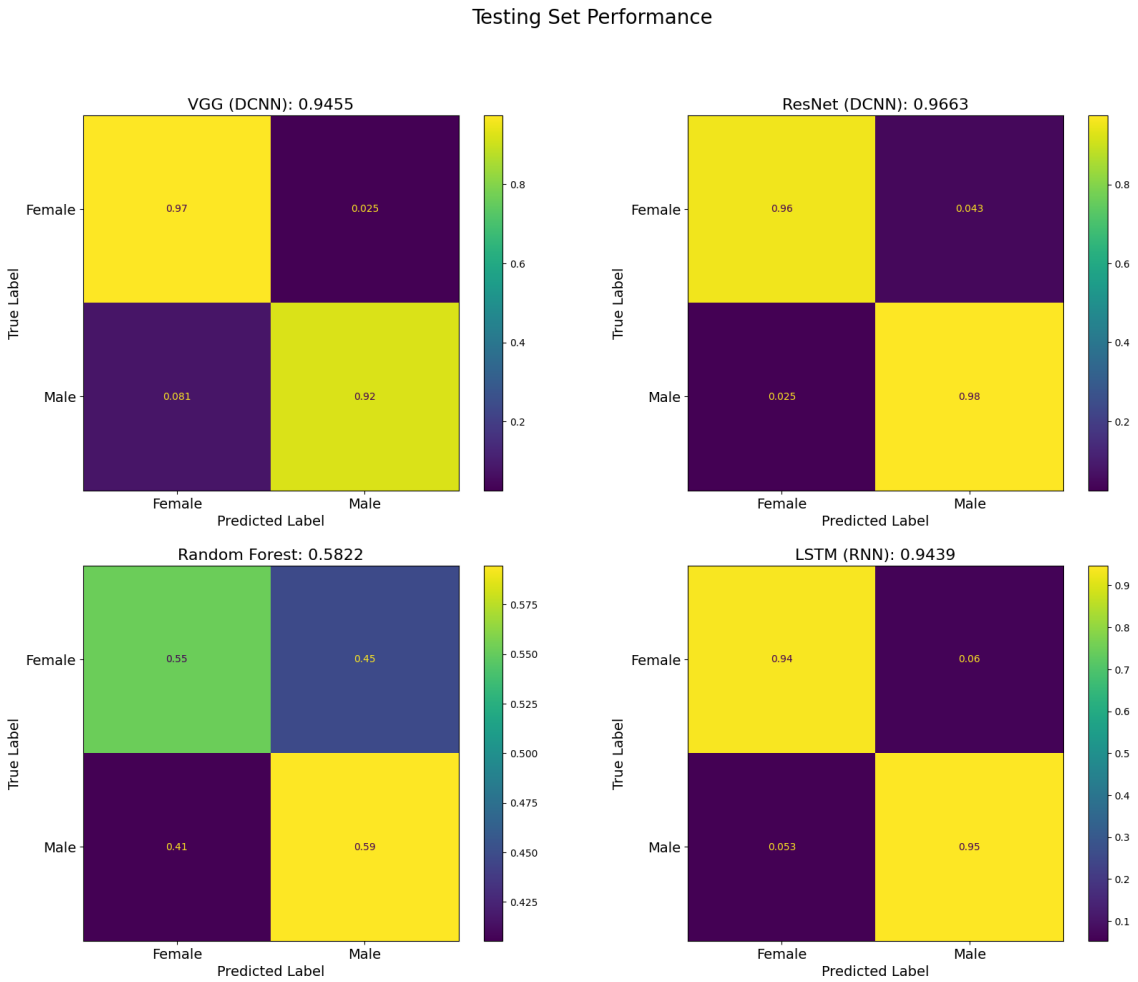

Figure 8: Confusion Matrix of all 4 models (Testing Set)

References

[1] Abeer Ali Alnuaim, Mohammed Zakariah, Chitra Shashidhar, et al. *Speaker Gender Recognition Based on Deep Neural Networks and ResNet50*. 2022. URL: https://www.hindawi.com/journals/wcmc/2022/4444388.

[2] Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim. *ACOUSTIC SCENE CLASSIFICATION USING PARALLEL COMBINATION OF LSTM AND CNN*. 2016. URL: https://dcase.community/documents/workshop2016/proceedings/Bae-DCASE2016workshop.pdf.

[3] Long Xu, Yi-Hua Yan, Xue-Xin Yu, et al. "LSTM neural network for solar radio spectrum classification". In: *Research in Astronomy and Astrophysics* 19.9 (2019). DOI: 10.1088/1674-4527/19/9/135. URL: https://iopscience.iop.org/article/10.1088/1674-4527/19/9/135/meta.

[4] Yan Zhang and Dan-jv LV. "Selected Features for Classifying Environmental Audio Data with Random Forest". In: *The Open Automation and Control Systems Journal* 7 (2015), pp. 135–142. URL: https://benthamopen.com/contents/pdf/TOAUTOJ/TOAUTOCJ-7-135.pdf.

[5] Anna Bosch, Andrew Zisserman, and Xavier Munoz. "Image Classification using Random Forests and Ferns". In: *2007 IEEE 11th International Conference on Computer Vision*. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4409066.

[6] B. Vimal, Muthyam Surya, Darshan, et al. "MFCC Based Audio Classification Using Machine Learning". In: *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 2021, pp. 1–4. DOI: 10.1109/ICCCNT51525.2021.9579881.

[7] Shaohua Wan, Yan Liang, and Yin Zhang. "Deep convolutional neural networks for diabetic retinopathy detection by image classification". In: *Computers Electrical Engineering* 72 (2018), pp. 274–282. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2018.07.042. URL: https://www.sciencedirect.com/science/article/pii/S0045790618302556.

[8] Vincenzo Dentamaro, Paolo Giglio, Donato Impedovo, et al. "AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath". In: *Pattern Recognition* 127 (2022), p. 108656. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2022.108656. URL: https://www.sciencedirect.com/science/article/pii/S0031320322001376.

[9] Amit Meghanani, Anoop C. S., and A. G. Ramakrishnan. "An Exploration of Log-Mel Spectrogram and MFCC Features for Alzheimer's Dementia Recognition from Spontaneous Speech". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. 2021, pp. 670–677. DOI: 10.1109/SLT48900.2021.9383491.

[10] Shengyun Wei1, Shun Zou1, Feifan Liao1, et al. "A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification". In: *Journal of Physics: Conference Series* (2020). DOI: 10.1088/1742-6596/1453/1/012085. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012085/pdf.

[11] Yasser Zeinali and Seyed Taghi Akhavan Niaki. "Heart sound classification using signal processing and machine learning algorithms". In: *Machine Learning with Applications* 7 (2022), p. 100206. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2021.100206. URL: https://www.sciencedirect.com/science/article/pii/S2666827021001031.

[12] *Anti-aliasing Filter Design and Applications in Sampling*. URL: https://resources.pcb.cadence.com/blog/2020-anti-aliasing-filter-design-and-applications-in-sampling.

[13] Karol J. Piczak. "Environmental sound classification with convolutional neural networks". In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2015, pp. 1–6. DOI: 10.1109/MLSP.2015.7324337.

[14] Shengyun Wei, Shun Zou, Feifan Liao, et al. "A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification". In: *Journal of Physics: Conference Series* (2019). eprint: https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012085/pdf.

[15] Leland Roberts. "Understanding the Mel Spectrogram". In: (). URL: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.

[16] Fatih Ertam. "An effective gender recognition approach using voice data via deeper LSTM networks". In: *Applied Acoustics* 156 (2019), pp. 351–358. ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2019.07.033. URL: https://www.sciencedirect.com/science/article/pii/S0003682X19304281.

[17] Tanvi Puri, Mukesh Soni, Gaurav Dhiman, et al. "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network". In: *Journal of Healthcare Engineering* 2022 (Feb. 2022), pp. 1–9. DOI: 10.1155/2022/8472947.

[18] Alexis Conneau Alexei Baevski. *Wav2vec 2.0: Learning the structure of speech from raw audio*. URL: https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/. (accessed: 06/06/2023).

[19] Dillon E. King. "The Inclusion of Sex and Gender Beyond the Binary in Toxicology". In: *Front Toxicol* 4 (July 2022). DOI: 10.3389/ftox.2022.929219. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9355551/.

[20] Anvarjon Tursunov, Mustaqeem, Joon Yeon Choeh, et al. "Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms". In: (2021). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8434188/.