# Speech-Based Models for Predicting Gender and Depression

Kalyani Jaware
University of Colorado Boulder

Dayn Reoh
University of Colorado Boulder

Erin Richardson
University of Colorado Boulder

Dhanavikram Sekar
University of Colorado Boulder

## 1 INTRODUCTION

Digital healthcare is increasingly leveraging speech-based ML technologies due to the ease of unobtrusive data collection through smartphones and wearables. Speech carries valuable insights into human behavior and mental state, stemming from the complex interplay of cognitive planning and articulation. Acoustic measures from speech, like prosody, provide critical information for mental healthcare, but may also be confounded by demographic factors, potentially leading to biases in ML algorithms detecting depression.

In this project we **build predictive models of depression and gender**. We also **explore gender bias** in these models. We use the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WoZ) dataset [1]. Further information about the dataset is displayed in Table 1.

| Category | Male | Female | Total |
|---|---|---|---|
| Train | 63 | 44 | 87 |
| Test | 14 | 17 | 20 |
| Total | 77 | 61 | 107 |

**Table 1: Train Test Split**

## 2 DATA PREPROCESSING

It was important to prepare and clean the dataset before beginning modeling efforts. The data cleaning process is outlined in Figure 1.

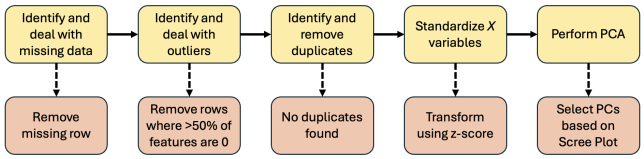


**Figure 1: Flowchart of Data Cleaning Process**

First, we concatenated the CSVs from different subjects to generate one large data matrix **X**. Next, we looked for missing data, finding and removing one row `NaNs`. When looking for outliers, we noticed that some rows had large amounts of features set to zero, including some features where a value of zero didn't make physical sense. It seemed that zero was used as an indicator of a missing value. To mitigate this (while trying not to remove too many useful observations), we removed any rows where over 50% of the feature values were zero. Next, we looked for duplicates but did not identify any duplicate rows. Finally, we standardized the variables by performing a z-score transformation to each feature column (but not to the response variables), ensuring that the features were on comparable scales.

We explored the use of Principal Components Analysis (PCA) for dimensionality reduction. We performed PCA and generated a Scree Plot to identify the appropriate amount of PCs to retain, deciding on 20 PCs. Ultimately, PCA did not improve our modeling performance so the PCA-transformed data was discarded.

## 3 DEPRESSION CLASSIFICATION

In the context of depression classification, the utilization of traditional machine learning algorithms, notably Logistic Regression and Random Forests, was undertaken with a subsequent comparison of their performance. To address the challenge posed by class imbalance, class weights were utilized for logistic regression, while a balanced modification of random forests was implemented. This modification ensured equitable representation of minority classes within each tree through a sampling strategy. The outcomes of logistic regression and Balanced Random Forests were similar in terms of balanced accuracy and Equality of Opportunity (EO) score.

Simultaneously, deep learning methodologies incorporating Convolutional Neural Networks (CNNs), LSTM-based recurrent neural networks, and Multi-Layer Perceptrons (MLPs) were explored to account for temporal dependencies within the data. Notably, MLPs emerged as the optimal choice, demonstrating the highest balanced accuracy and EO score among the examined deep learning architectures.

The results of these analyses is encapsulated in 2, presenting a comprehensive overview of the comparative performance metrics of the best-performing aforementioned algorithms in the context of depression classification.

| Model | Acc. | Bal. Acc. | EO |
|---|---|---|---|
| Logistic Regression | 0.60 | 0.57 | 0.40 |
| Balanced Random Forests | 0.60 | 0.57 | 0.40 |
| Multi-Layer Perceptrons | 0.80 | 0.81 | 0.80 |
| Convolutional Neural Networks | 0.60 | 0.47 | 0.80 |

**Table 2: Comparison of model performance metrics for Depression classification**

From the results, it is evident that weighted Multi-layer perceptrons were far superior in identifying depression compared to other algorithms. Only sample in depression class was misclassified.

## 4 GENDER CLASSIFICATION

Initially, the top-performing algorithms mentioned in previous sections were employed for gender classification. Upon model refinement, it was discerned that gender can be classified with remarkable precision. Table 3 contains a summary of the results.

| Model | Accuracy | Balanced Accuracy |
|---|---|---|
| Logistic Regression | 1.00 | 1.00 |
| Random Forests | 1.00 | 1.00 |
| Multi-Layer Perceptrons | 1.00 | 1.00 |

**Table 3: Comparison of model performance metrics for Gender classification**

Based on the results, it can be seen that all the models - Logistic Regression, Multi-layer Perceptrons, and Random Forests were able to classify gender with remarkable 100% accuracy.

To make sure that the models are not overfitting and to interpret the results, a simple decision stump was constructed. This root node for the stump was selected based on the feature importances extracted from the Random Forests model. It was observed that the feature 'F0semitoneFrom27.5Hz_sma3nz_percentile50.0', which represents the 50th percentile of logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0) or in other words the feature indicative of voice pitch, alone could classify gender with 95% accuracy. However, employing an additional six features enabled achieving 100% accuracy in gender classification. Figure 2 contains the visualization of the decision tree.

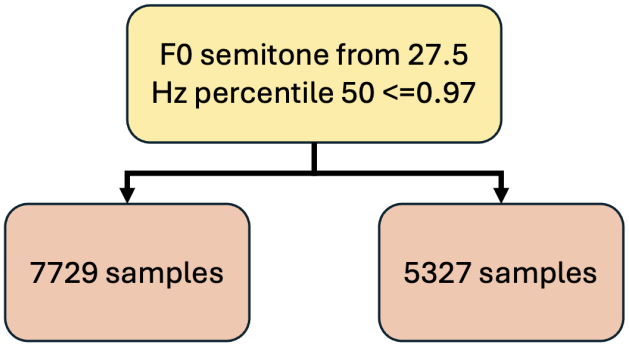


**Figure 2: Decision tree for gender classification (Depth 1)**

## 5 IDENTIFYING INFORMATIVE FEATURES

### 5.1 Feature Selection for Depression using Pearson's Correlation

Traditional filter methods, such as Pearson's correlation, were employed to identify a specific subset of features highly correlated with depression while being uncorrelated with each other. Pearson's correlation uses the covariance between the features and their individual standard deviation to find out if they have any relationship between them.

The top 20 features based on Pearson's correlation are selected and their scores are displayed in figures 3 and 4. Figure 3 and 4 represents the top 10 positively correlated features and the top 10 negatively correlated features respectively.

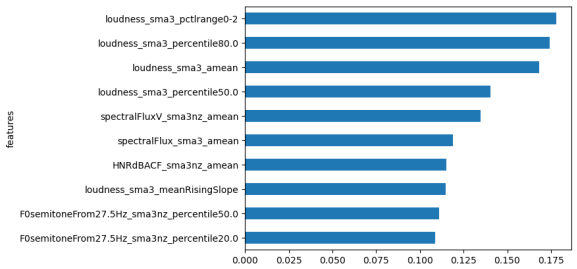


**Figure 3: Top 10 features positively correlated with Depression based on Pearson's Correlation**

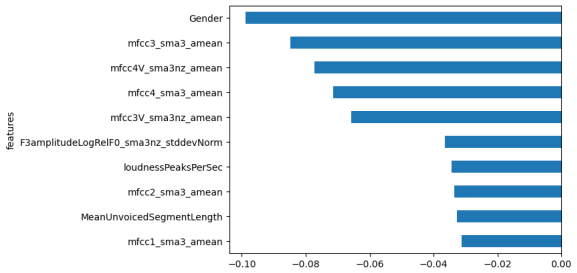


**Figure 4: Top 10 features negatively correlated with Depression based on Pearson's Correlation**

The most interesting finding from figure 4 was the Gender was the feature that has the highest negative correlation with Depression. But since the correlation value is very weak(-0.1), this cannot be considered as strong correlation. None of the features had a correlation value more than 0.2. Which means all the features are somewhat weakly correlated with depression.

Similarly, The subset size ranged from 5 to 85, incremented by 5. The subset of features comprises of both positively correlated and negatively correlated features. The same Multilayer Perceptron (MLP) model, which demonstrated optimal performance in depression classification, was utilized. The model's performance across different subset sizes is depicted in Figure 5.

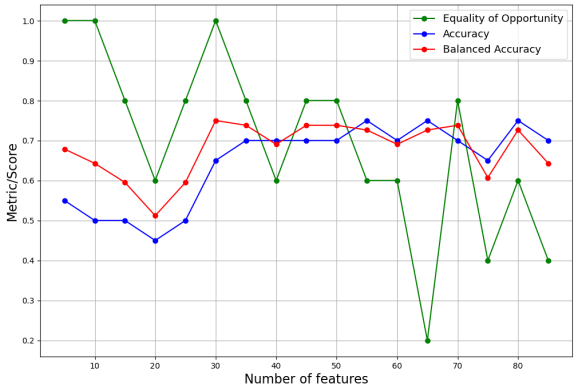


**Figure 5: Feature Selection for Depression using Pearson's Correlation**

Figure 5 illustrates the outcomes of feature selection for depression utilizing Pearson's correlation. It is evident that the model achieves its peak performance when 30 features are used. On increasing the number of features, the accuracy and balanced accuracy remains almost the same while the EO score decreases. This performance achieved using 30 features aligns closely with the performance achieved using the entire feature set.

## 5.2 Feature Selection for Gender using Pearson's Correlation

For gender classification, a similar methodology was applied. The 20 most informative features regarding gender were identified, with the top 10 positively correlated features depicted in Figure 6 and the top 10 negatively correlated features in Figure 7.

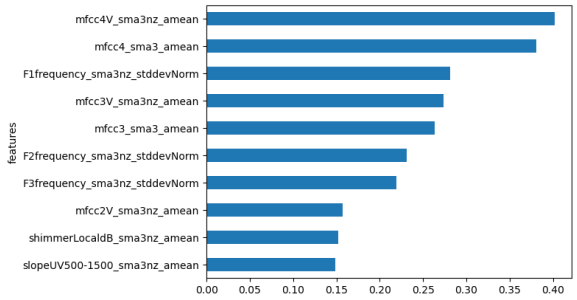


**Figure 6: Top 10 features positively correlated with Gender based on Pearson's Correlation**

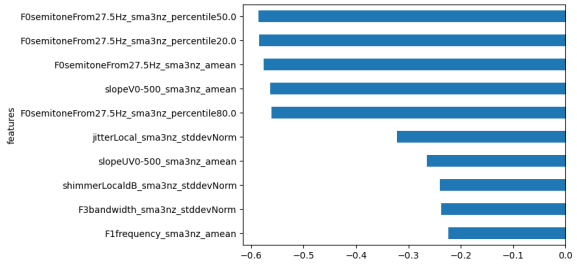


**Figure 7: Top 10 features negatively correlated with Gender based on Pearson's Correlation**

Figure 6 showcases features positively correlated with gender, predominantly associated with speech signal frequency. Conversely, Figure 7 displays features negatively correlated with gender, primarily indicative of voice pitch, which emerged as the most informative for gender classification.

Utilizing the same MLP model employed in the preceding analysis, gender classification was conducted using these features. The n features chosen comprises the top n/2 positively correlated and top n/2 negatively correlated features. The outcomes are presented in Figure 8.

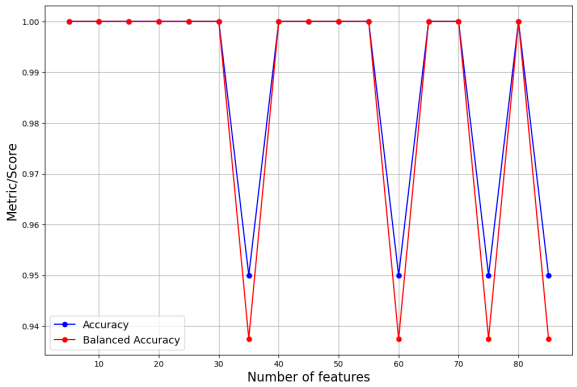


**Figure 8: Feature Selection for Gender using Pearson's Correlation**

The results illustrate that merely the top 5 features correlated with gender are adequate to achieve 100 percent accuracy in gender identification. This finding complements the results obtained through feature importance-based selection, underscoring the significance of pitch-based features in gender characterization.

## 5.3 Feature Selection for Depression using Fisher's criterion

The other filter method that was experimented with is the Fisher's criterion. This method aims to maximize the ratio of the variance between classes to the variance within classes. Each feature has a Fishers Score associated with depression classification, and the highest score represents the features that provide the best separation between classes. Fisher's criterion rendered higher average accuracy and balanced accuracy compared to Pearson's Correlation.

Utilizing the same MLP model as a benchmark, various feature subsets ranging in size from 5 to 85, with an incremental step size of 5, were supplied to the model. Subsequently, performance metrics were recorded and analyzed. The findings are presented in Figure 9 for depression classification and Figure 10 for gender classification.

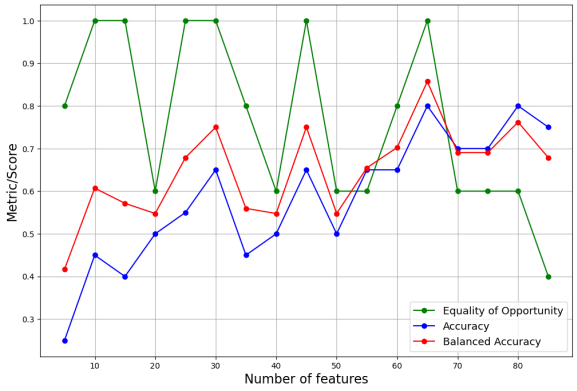


**Figure 9: Feature Selection for Depression using Fisher's Criterion**

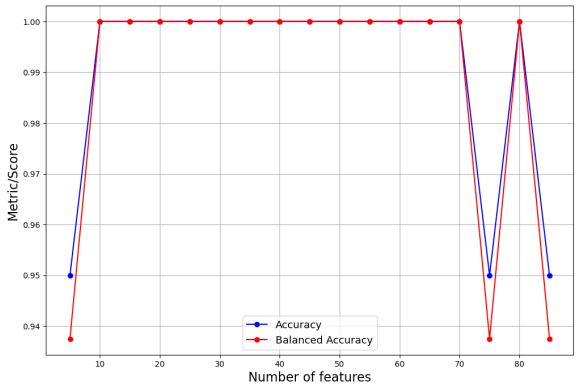


**Figure 10: Feature Selection for Gender using Fisher's Criterion**

Based on the results, Fisher's criterion proves effective for feature selection in both gender and depression classification. For gender classification, just 10 features achieve 100 % accuracy, demonstrating efficiency in pinpointing critical discriminative features. In depression classification, using about 30 features optimizes balanced accuracy, indicating that beyond this, additional features add little value.

## 5.4 Feature Selection using Feature Importance

In the initial filtering approach, the feature importance derived from the random forest model was employed to identify the most informative features for both gender and depression classification. Through experimentation with the different number of features ranging from 10 to 80, it was observed that utilizing the 60 most significant features enabled the construction of a random forest model comprising 300 trees, each with a depth of 200. This model achieved a balanced accuracy of 60% in classifying depression, along with an Equality of Opportunity (EO) score of 0.6. Figure 11 shows the evaluation metrics of the model for different number of features.

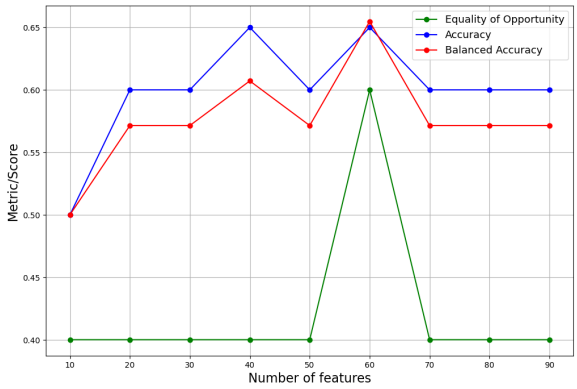


**Figure 11: Feature Selection for Depression using feature importance**

Similarly, in gender classification, employing just one feature allowed the model to achieve a notable accuracy rate of 95%. This was also discussed in earlier sections. However, attaining a perfect classification accuracy necessitated using seven features. On analyzing this, it was found that these seven features that represent the pitch and amplitude of a person's voice. The performance metrics associated with the increasing number of features utilized is depicted in 12.

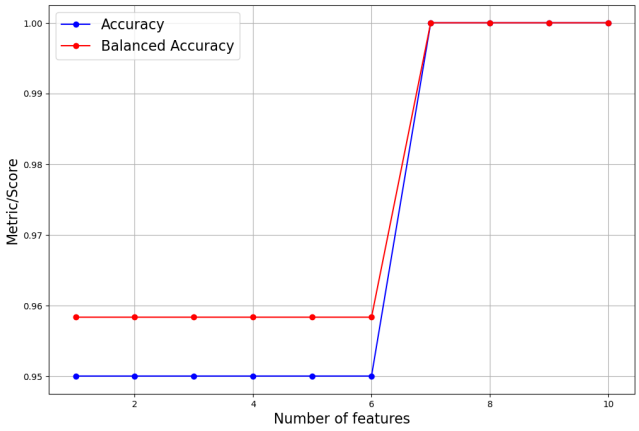


**Figure 12: Feature Selection for Gender using feature importance**

# 6 BIAS MITIGATION

## 6.1 Removal of Gender Dependent Features

In this analysis, eight gender-dependent features were excluded from the original dataset. The identification of these gender-dependent features was guided by a feature importance analysis conducted using a random forest model. The top-performing model from prior trials - the MLP model with class weights, was applied. Despite a marginal accuracy dip, there was a notable reduction in the equal opportunity (EO) score metric. This shift suggests that the exclusion of informative gender features not only impacted overall accuracy but also led to a less fair classification process. The results of this experiment are depicted in Table 6.

| Metric | Result |
|---|---|
| Accuracy | 0.75 |
| Balanced Accuracy | 0.73 |
| Equality of Opportunity | 0.6 |

**Table 4: Evaluation metrics (Features removed based on Feature Importance)**

To investigate the reduction in performance, a confusion matrix was plotted. It can be seen that the mode performance on an average level in both Depression and Non-Depression classes. The results can be seen in 13

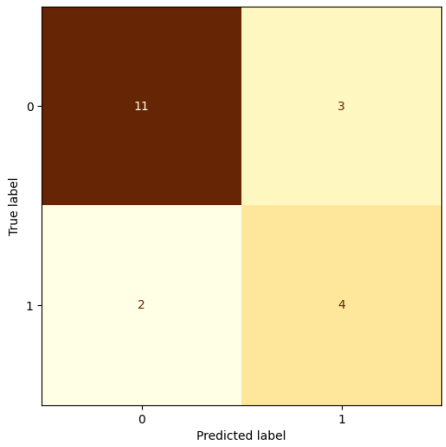


**Figure 13: Results after Bias Mitigation by eliminating gender dependent features**

Furthermore, another attempt was made to exclude the 5 most gender informative features chosen based on Pearson's correlation metric. It was observed that both Balanced Accuracy an EO scores decreased significantly. The results are shown in Table 5. The confusion matrix of the resultant model's performance on test set in shown in Figure 14.

| Metric | Result |
|---|---|
| Accuracy | 0.6 |
| Balanced Accuracy | 0.57 |
| Equality of Opportunity | 0.4 |

**Table 5: Evaluation metrics (Features removed based on Pearson's Correlation)**

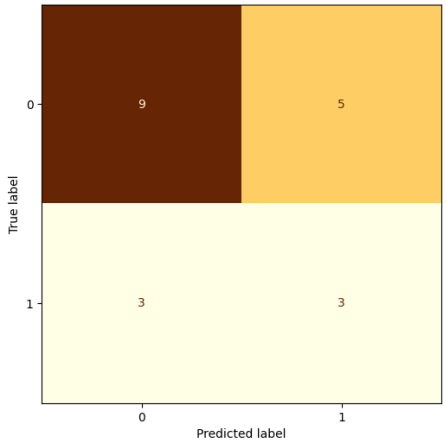


**Figure 14: Results after Bias Mitigation by eliminating gender dependent features based on Pearson's Correlation**

This emphasizes a need for thoughtful feature selection to mitigate bias and promote fairness in model selection in case of real-world applications.

## 6.2 Experimental approaches to mitigate bias

This experiment attempted to mitigate gender bias in depression prediction by training a neural network model (MLP) that simultaneously predicts gender and depression. Class weights have also been used for depression classification in order to handle the class imbalance problem. This approach aims to help the model learn the most informative features for gender as well as depression concurrently which ultimately helps in mitigating the bias. With this method, the accuracy dropped significantly but the EO score was better than the gender dependent feature elimination method. The results are shown below.

| Metric | Result |
|---|---|
| Accuracy | 0.55 |
| Balanced Accuracy | 0.44 |
| Equality of Opportunity | 0.8 |

**Table 6: Evaluation metrics (Simultaneous Prediction using MLP)**

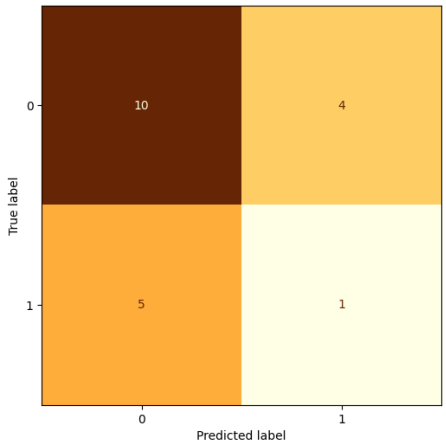


**Figure 15: Confusion Matrix - Bias Mitigation through simultaneous prediction**

# 7 FUTURE WORK

Going forward, this project could be improved in several ways:

(1) Internal validation should be implemented to improve the reliability of model performance estimates. Currently, model performance was evaluated on the provided held-out test set. Cross-validation or bootstrap resampling could be used to generate multiple performance estimates that could be aggregated, leading to final performance estimates robust to the noise in estimates generated on individual iterations.

(2) Ensemble learning should be implemented to leverage the strengths of different machine learning models and ultimate generate stronger final predictions. In this project, we found different machine learning model types (such as random forests and neural nets) to have their own advantages and disadvantages. Ensemble methods (such as boosting algorithms) can aggregate predictions from multiple weak learners (that have different strengths and weaknesses) to provide stronger overall final estimates.
(3) Bias from a broad range of sources (beyond gender) should be investigated. Our models may be biased along dimensions other than gender, such as race, first language, etc., and its outcomes may affect subgroups unjustly. We should look to understand the equality of opportunity provided to subgroups across multiple demographic dimensions. We should also work to interpret our models' predictions in the context of race- or language-sensitive features, for example.

## 8 CONCLUSION

In this project we developed speech-based models for predicting depression with a focus on model fairness with respect to gender. Along we way, we developed models to classify gender and found features important to each of gender and depression. By removing gender-sensitive features from the candidate feature set, we were able to improve the Equality of Opportunity of our depression models with only a small drop in overall accuracy.

## REFERENCES

[1] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland, 3123–3128. http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf