# Mitigating socio-demographic bias in language-based machine learning models of depression

Harish Nandhan Shanmugam
Manidatta Anumandla
Manohar Korikana
Ajay Tata

# Goal and Methodology

Goal is to estimate the depression severity (PHQ-8 score) from participants speech transcripts.
And Address bias in predictions across the gender and race/ethnicity groups.

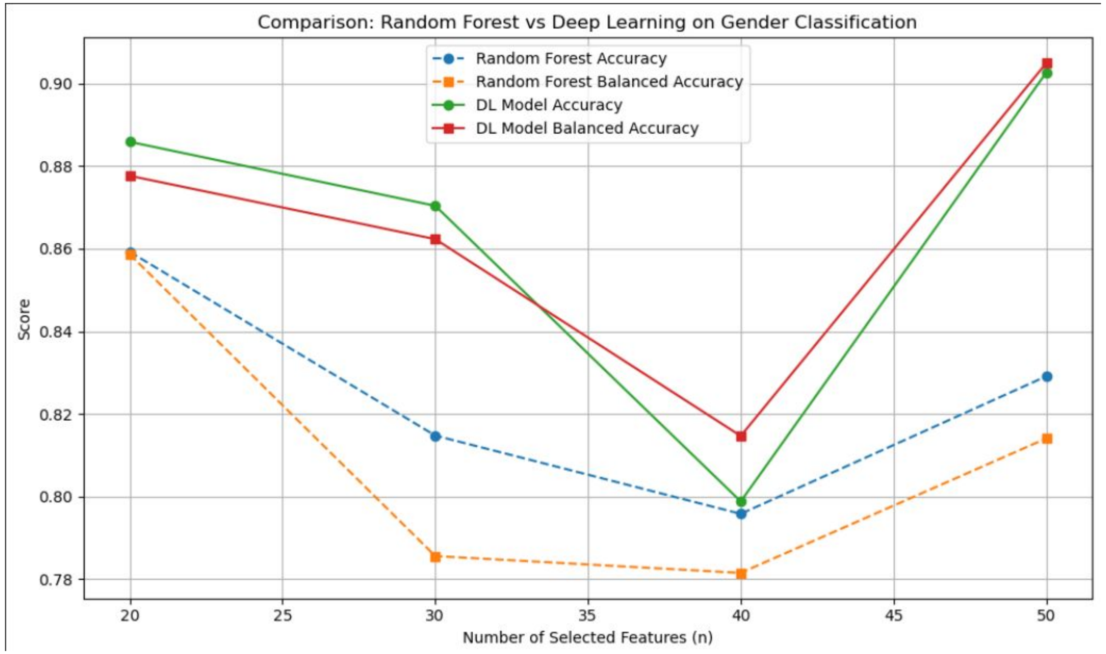
**Dataset:**

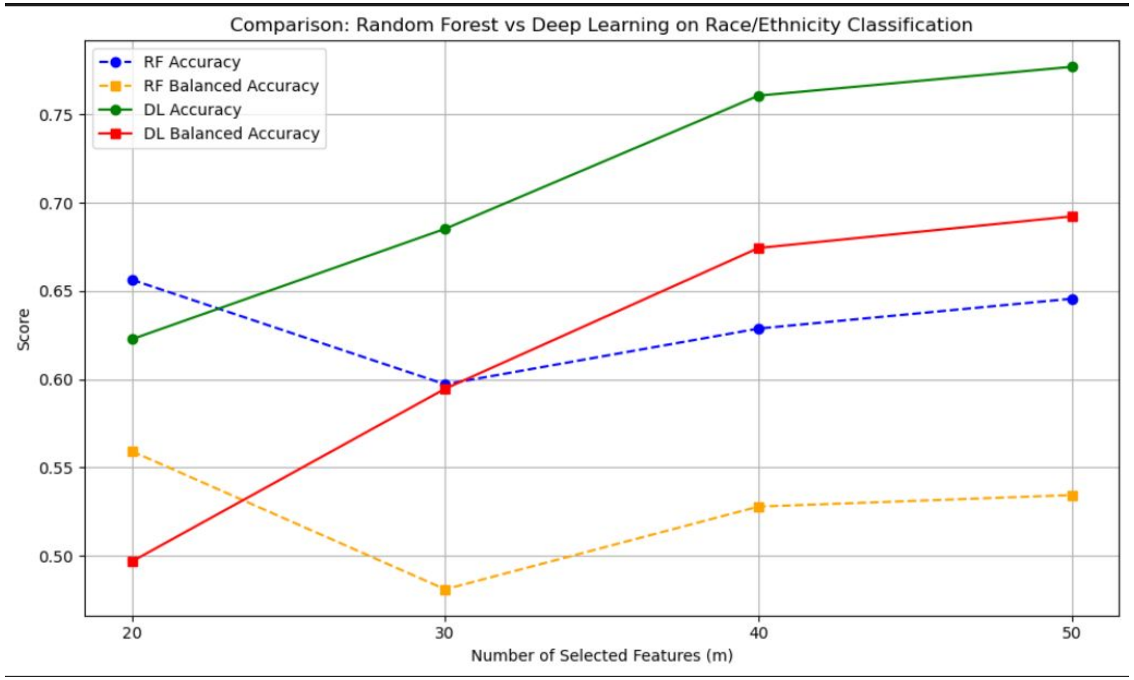
- Transcripts data: Contains text data of the patients
- DAIC demographic data: Contains the data of patient like Gender, Race, PHQ-score

**Methodology:**

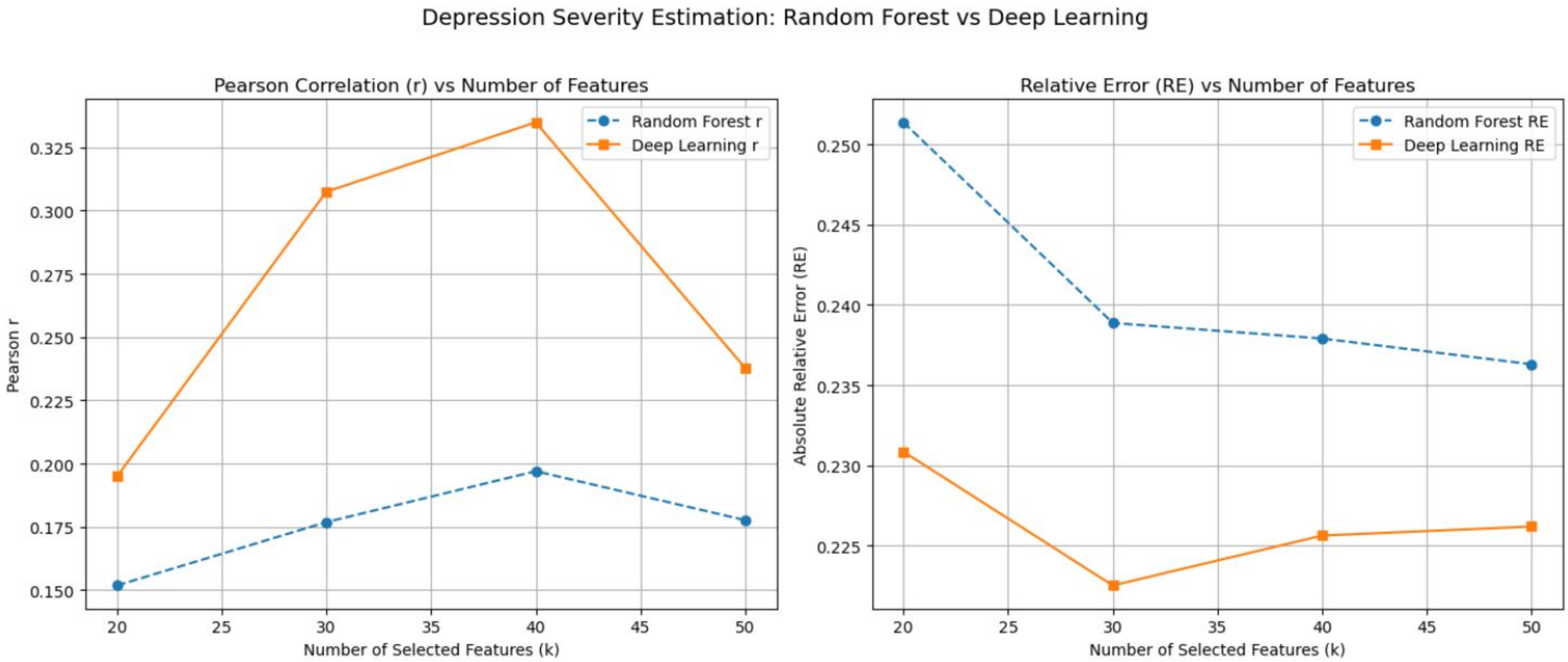
- Cleaning the transcripts and encoding the demographic attributes.
- TF-IDF vectorizer for syntactic, Vader for Semantic.
- Feature selection: selecting the top informative features for prediction.
- Model Training: Random Forest, Deep Learning, GPT-2 with few-shot learning.
- Evaluation: Accuracy, Balanced Classification Accuracy, Pearson Correlation Coefficient(r), Absolute Relative Error(re), Group-wise fairness breakdown.

# Gender and Race Classification Results:



# Depression Severity Estimation Results:

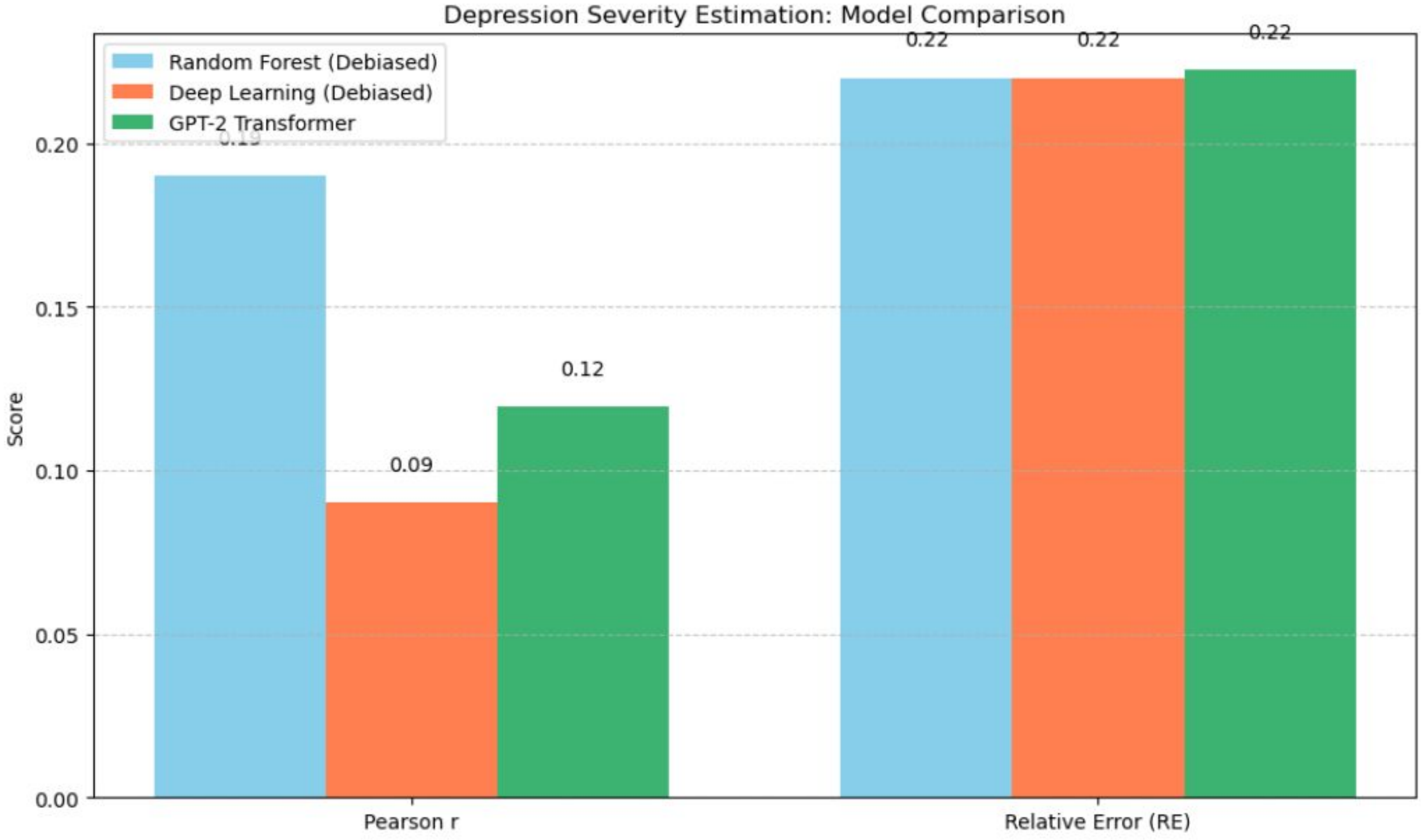


# Depression Severity Estimation Results based on the Gender - Ethnicity group:

```
Groupwise Results (Random Forest):
                        Group  Pearson_r        RE
0    Female - White American   0.476472  0.291818
1    Male - African American   0.968791  0.264821
2            Male - Hispanic   0.989584  0.432157
3      Male - White American  -0.144179  1.520000
4  Female - African American   0.957575  0.263889

Groupwise Results (Deep Learning):
                        Group  Pearson_r        RE
0    Female - White American   0.492863  0.375104
1    Male - African American   0.629065  0.299063
2            Male - Hispanic   0.831736  0.623356
3      Male - White American  -0.444475  0.385798
4  Female - African American   0.935017  0.298812
```

# Depression Severity Estimation Results for the Debiased Model:




```
[RF] Female - White American → r: 0.9827 | RE: 0.0738
[RF] Male - African American → r: 0.7489 | RE: 0.1390
[RF] Male - Hispanic → r: 0.8653 | RE: 0.1260
[RF] Male - White American → r: 0.9146 | RE: 0.1061
[RF] Female - Hispanic → r: 0.8905 | RE: 0.1703
[RF] Female - African American → r: 0.8594 | RE: 0.1470
```

```
Female - White American → r: 0.1454 | RE: 0.2297
Male - African American → r: 0.2035 | RE: 0.2096
Male - Hispanic → r: nan | RE: 0.2314
Male - White American → r: -0.0047 | RE: 0.2099
Female - Hispanic → r: 0.5587 | RE: 0.3968
Female - African American → r: -0.1705 | RE: 0.2583
```

# Conclusion:

**Gender and Race Classification:**

- Random Forest seems sensitivity to feature quantity - too few or too many affects it.
- Deep Learning although briefly inconsistent, ends up generalizing better at higher feature counts.

**Depression severity Estimation:**

- Deep Learning shows more predictive power, especially in how well it tracks the actual PHQ scores.
- Random Forest performs steadily but cannot match deep learning, especially in correlation.

**Depression Severity Estimation for Debiased Model:**

- After debiasing, all models converge to a similar error. They all generalize similarly when demographic shortcuts are removed.
- However GPT-2 maintains a slight edge in correlation.
- Deep Learning is impacted more heavily in terms of correlation but stable in average prediction error.
- Random Forest, though it is more interpretable, performs similarly in RE but doesn't capture complex complex linguistic signals.
- Over-all, GPT-2 remains promising , even in debiased settings, and could benefit from further fine-tuning or domain-specific prompts.