Characterizing Bias in Machine Learning Algorithms for Detecting Alzheimer's Disease

Priya Bhirgoo[1], Sierra Gong[1], Anya Greenberg[2], Nidhish Nerur[1]

1. Massachusetts Institute of Technology, Cambridge, MA 2. Harvard University, Cambridge, MA

Introduction

- Alzheimer's disease (AD) is the leading cause of dementia, posing growing clinical, economic, and societal challenges[1]
- Early detection is critical for slowing disease progression, yet current diagnostic approaches rely heavily on specialists and biomarker tests, limiting accessibility[1]
- In this study, we aim to develop an ML model that detects AD from clinical data while minimizing bias, ensuring equitable performance across patient groups

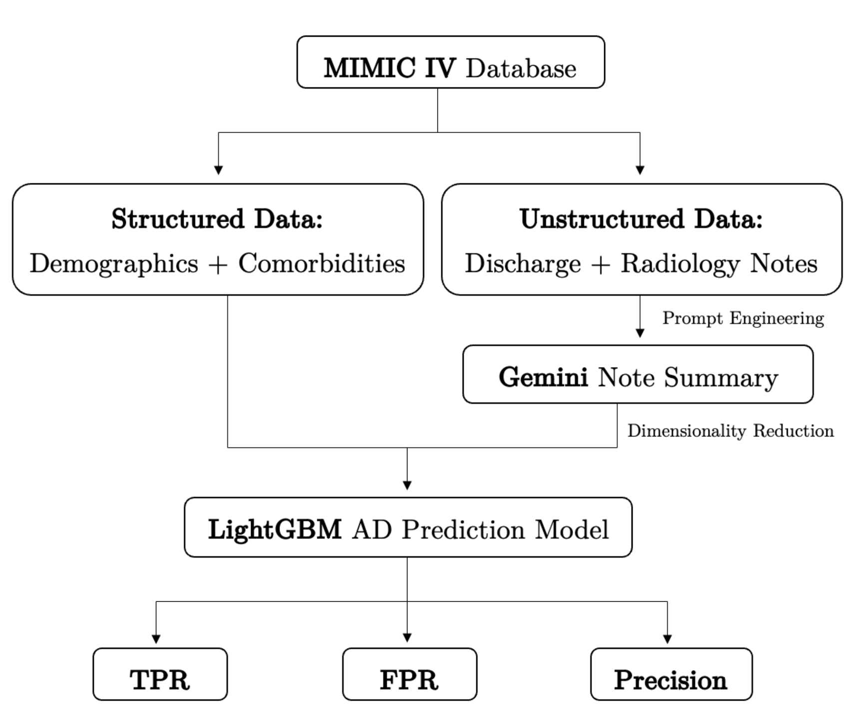
Data Preparation & Cohort Building

- We utilized **demographic and comorbidity data** as well as **discharge and radiology notes** from the MIMIC IV dataset
- Derived features were created to group subcategories, address missing values, and reduce sparsity across features
- Case cohort was built using ICD-10 codes representing AD diagnosis; control cohort was built excluding ICD-10 codes representing ADRD diagnosis
- **Propensity score matching** using a logistic regression model was applied to identify controls with complete records which were demographically similar to the AD subjects in the case cohort

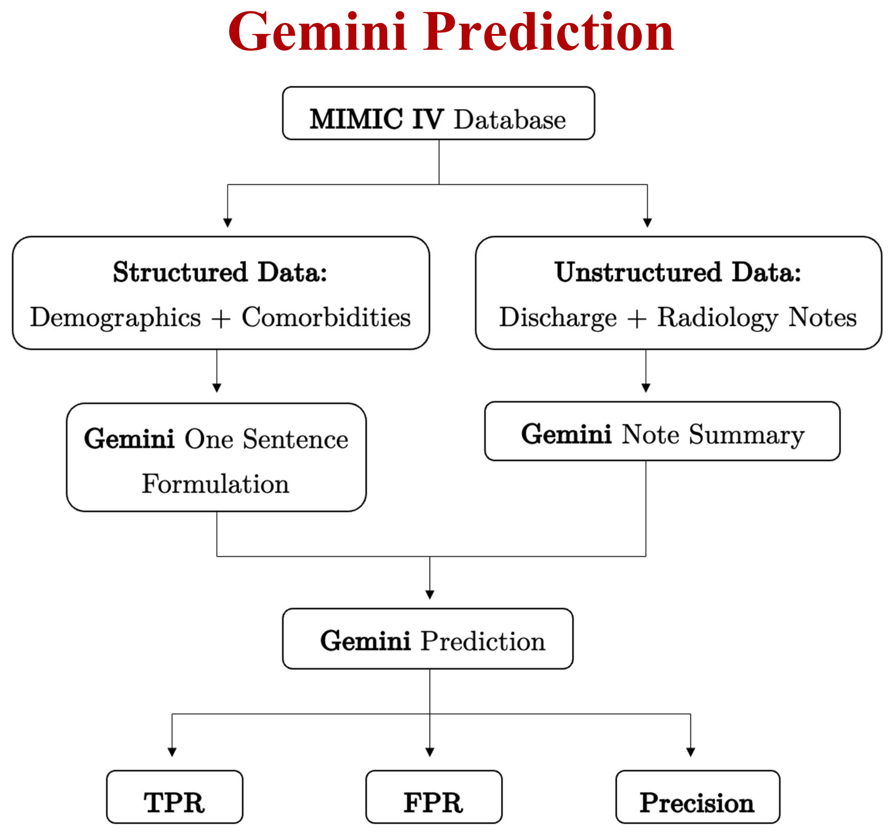
Category	# Cases	# Controls
Total Count	1092	1084
Age		
Mean	81.655	81.146
Standard Deviation	8.938	9.083
Median	83	83
Min	48	47
Max	100	100
Self-reported Gender		
Female	685	672
Male	407	412
Self-reported Racial Group		
African America	135	138
Asian	33	30
Hispanic/Latino	50	45
White	786	779
Other	92	88
Language		
English	914	917
Non-English	178	167
Insurance Source		
Medicaid	84	69
Medicare	943	947
Other	65	68

Methodology

LightGBM Prediction Model



Gemini Prediction



Three key metrics were evaluated to assess model fairness:

1. **Equal Opportunity:** measured by TPR
2. **Predictive Parity:** measured by Precision
3. **Equalized Odds:** evaluated as the joint parity of TPR and FPR

For Equal Opportunity and Predictive Parity, two statistical tests were applied:

1. Subgroup-pair level: **Two-proportion Z-test** to detect significant differences between each subgroup pair within demographic attributes

2. Overall group level: **Chi-squared test** to determine if TPR or Precision significantly varied across all subgroups of a demographic attribute

All tests included bootstrap procedures to estimate 95% confidence intervals. Bias detection thresholds followed industry guidelines:

Equal Opportunity & Predictive Parity: $p < 0.05$

Equalized Odds Difference : > 0.1

Results

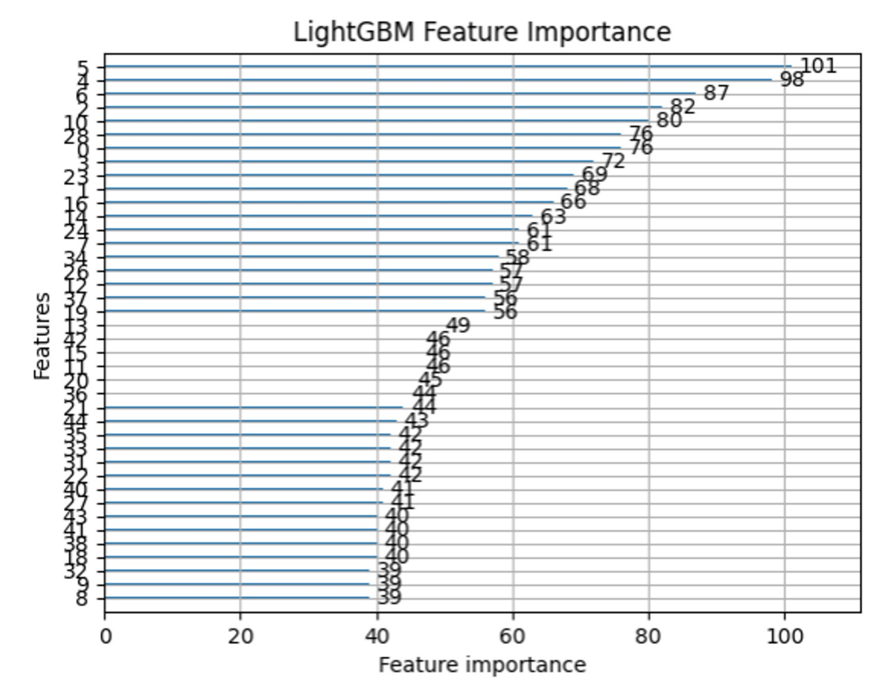
LightGBM Prediction Model

Training Set Performance

Metric	Value
True Positive Rate (TPR)	0.843
False Positive Rate (FPR)	0.112
Precision	0.882
Sample Size	1740

Test Set Performance

Metric	Value
True Positive Rate (TPR)	0.648
False Positive Rate (FPR)	0.325
Precision	0.684
Sample Size	436



Race Group	Equal Opportunity Range	FPR Range	Predictive Parity Range	Sample Size
Training Set				
White	0.814–0.865	0.088–0.132	0.857–0.902	1267
African American	0.739–0.879	0.038–0.129	0.876–0.961	210
Hispanic/Latino	0.685–0.900	0.000–0.163	0.790–1.000	75
Other	0.779–0.927	0.138–0.328	0.724–0.890	136
Asian	1.000–1.000	0.000–0.194	0.852–1.000	50
Testing Set				
White	0.570–0.701	0.288–0.430	0.605–0.734	327
African American	0.572–0.871	0.062–0.316	0.622–0.918	54
Hispanic/Latino	0.143–0.714	0.000–1.000	0.250–1.000	15
Other	0.516–0.970	0.000–0.372	0.500–1.000	27
Asian	0.500–1.000	0.000–0.647	0.263–1.000	13

- Chi-squared tests did not reveal any statistically different differences across racial groups
- Pairwise comparisons were not statistically different, however we have small sample sizes and wide confidence intervals

Gemini Prediction

Overall Model Performance

Metric	Value
True Positive Rate (TPR)	0.634
False Positive Rate (FPR)	0.241
Precision	0.726
Sample Size	2176

Demographic Group	Equal Opportunity	Predictive Parity	Equalized Odds
Age Group	✗	✗	✗
Sex	✓	✓	✓
Insurance Group	✓	✓	✗
Language Group	✓	✓	✓
Race Group	✓	✓	✗

- Significant fairness disparities observed in TPR and precision across age groups, particularly poorer performance in individuals over age 90 ($p < 0.01$)
- Pairwise racial subgroup comparisons identified notable sensitivity differences, especially between African American and Hispanic/Latino individuals ($p = 0.0328$)
- Equalized Odds analyses showed substantial inequities in model performance between the youngest (<70 years) and oldest (>90 years) groups, exceeding established fairness thresholds

Conclusions

- LightGBM achieved comparable predictive performance to Gemini
- Despite similar accuracy, the Gemini-based approach showed significantly greater fairness disparities across age and racial groups
- Future work includes:
  1. broadening the scope of fairness evaluation using additional metrics such as demographic parity, calibration within groups, and treatment equality
  2. investigating bias mitigation strategies across the ML pipeline for the LightGBM approach
  3. examining the internal decision-making processes of the Gemini model

Acknowledgements/References

We would like to thank our project mentors, Dr. Sudeshna Das Yingnan He, and Yu Leng, as well as the 6.7930 teaching staff, Dr. Peter Szolovits, Dr. David Sontag, Sofie and Ilker without whom this study would not have been possible.

1. Scheltens, Philip et al. Alzheimer's disease. Lancet. 397,10284 (2021). https://doi:10.1016/S0140-6736(20)32205-4