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

Priya Bhirgoo, Sierra Gong, Anya Greenberg, Nidhish Nerur

Section 1: Introduction

Section 1.1: Clinical importance

The aim of this study is to develop and evaluate a machine learning (ML) model for detecting Alzheimer's disease (AD) using electronic health records. AD is the leading cause of dementia and represents a growing economic, clinical, and societal burden worldwide [1]. Diagnosis at an early stage is key to disease management through lifestyle and existing therapeutic interventions [1]. Currently, diagnosis is reliant on clinical evaluation and biomarker tests which are not readily available [1]. An ML model trained to evaluate and detect AD from non-specific clinical notes may enhance clinicians' ability to detect AD early and reduce economic, clinical, and societal burden over time. However, ML models applied in healthcare settings have been shown to exacerbate existing health inequities [2]. Therefore, it is important to evaluate bias and fairness in order to develop equitable ML models, which is what we aim to do in this study.

Section 1.2: Basic approach and conclusions

We utilized demographic and comorbidity data from MIMIC-IV (Medical Information Mart for Intensive Care IV) [3-5] as well as discharge summaries and radiology notes from MIMIC-IV-NOTE [3,4,6]. We selected patients with ICD-10 codes of AD and matched controls in a 1:1 ratio using a propensity score model based on logistic regression; these AD and control patients constituted our cohort. For the whole cohort we extracted relevant demographic and comorbidity data and summarized chronologically concatenated discharge summaries and radiology notes for each patient, making sure to mask any mention of AD. Using these masked and summarized notes as well as the demographic and comorbidity tabular data, we trained various detection models including multi-modal models using light gradient boosting machine (lightGBM) models, and direct prediction with Gemini, a large language model (LLM). We evaluated these models using established fairness metrics: equal opportunity, predictive parity, and equalized odds. We found that, overall, the multi-modal models with careful tuning performed comparably to the Gemini-based model in terms of predictive performance. Using the LightGBM model with our baseline prompt (Prompt 1), we achieved a true positive rate (TPR) of 0.648, a false positive rate (FPR) of 0.325, and a precision of 0.684. In comparison, the Gemini-based model achieved a TPR of 0.634, FPR of 0.241, and precision of 0.726. However, pairwise comparisons of fairness metrics revealed that the Gemini-based model exhibited significantly greater bias across age groups and certain racial groups, particularly African American and Hispanic/Latino populations. In contrast, the LightGBM model showed no statistically significant differences across any demographic groups in two out of the three fairness metrics evaluated.

Section 1.3: Key contributions

This work makes several key contributions to the development of fair and effective machine learning models for Alzheimer's disease (AD) detection. First, we developed models that use clinical notes to detect AD, enabling explicit evaluation and characterization of bias. Second, we conducted a comprehensive bias assessment for each model using established fairness metrics: equal opportunity, predictive parity, and equalized odds. Third, we evaluated the performance of multimodal models - those combining structured clinical data with unstructured notes - against a model based solely on Gemini, a large language model (LLM). This comparison provides insights into the trade-offs between model complexity and fairness in AD detection, supporting future model selection that allows equitable performance across diverse patient demographics.

Section 2: Related work

Approach	% of studies	Description
Rule-based	67%	Manual pattern matching, keyword extraction, clinical ontologies
Traditional ML	28%	Logistic regression, random forests trained on extracted features from notes
Deep Learning	17%	Transformer-based models like ClinicalBERT fine-tuned on clinical notes

Table 1: Summary of NLP Approaches Used in Studies for Detecting Cognitive Impairment

Shankar et al. (2025) conducted a literature review on the use of natural language processing (NLP) of health records for detection of cognitive impairment from clinical notes using rule-based algorithms, traditional ML, and deep learning. The NLP methodologies employed across the reviewed studies are summarized in Table 1. The authors found that 18 of these models were robust with median sensitivity 0.88 (IQR 0.74-0.91) and specificity 0.96 (IQR 0.81, 0.99) [7]. However, only a minority of studies assessed model performance across demographic subgroups such as race, self-reported sex, and education. The review also notes potential sources of algorithmic bias, including skewed training data, diagnostic disparities, and inconsistent documentation practices. None of the studies implemented formal bias mitigation techniques.

Furthermore, a recent study evaluated GPT-4o's ability to determine and differentiate stages of cognitive impairment based on unstructured clinical notes and achieved high kappas on the two tasks (0.83-0.96) [8]. This study demonstrated LLM's potential as a scalable chart review tool to assist in AD diagnosis. Although these models highlight the feasibility of using ML and LLMs with unstructured clinical notes to detect AD, there was limited analysis of sociodemographic bias. Because of the potential for dataset shift and privacy standards for clinical data, we aim to train our own model on MIMIC-IV data in order to investigate potential bias that may arise in healthcare applications of ML and LLMs.

Section 3: Data and Experiment Setup

Section 3.1: Data Acquisition and Preparation

Data for this study was obtained from the MIMIC-IV database, a large, publicly available repository of de-identified health information from patients admitted to the critical care units of Beth Israel Deaconess Medical Center. MIMIC-IV contains data collected between 2008 and 2019, primarily sourced from Metavision bedside monitoring systems [3-5]. The initial phase of this study involved the acquisition, preparation, and exploratory analysis of data from MIMIC-IV, with the goal of establishing a robust foundation for the development of an AI model to detect AD.

We used data from the following files: patients.csv.gz, admissions.csv.gz, diagnoses_icd.csv.gz, d_icd_procedures.csv.gz, prescriptions.csv.gz, discharge.csv.gz, radiology.csv.gz. Only patients that had records concerning diagnosis details, demographic information, and discharge summaries for at least one hospital admission were included. This ensures data quality, as incomplete records could compromise the analysis or introduce bias. For cases, the hospital admission ID corresponding to the patient's first AD diagnosis was selected. Additionally, derived features were created to complement the dataset by grouping subcategories, addressing missing values, and reducing sparsity across a number of features including self-reported race, language, and insurance provider type. The set of derived features in the dataset is further explained in Appendix A.

Discharge and radiology notes were extracted and appended to the dataset as they provide access to rich, unstructured clinical information that may contain early signs or subtle indicators of cognitive decline not captured in structured data alone. Discharge notes are long-form clinical narratives that summarize a patient's hospitalization. They typically describe the reason for admission, the key events, and treatments during the hospital stay (the

"hospital course"), and any important instructions or recommendations for the patient after discharge [4,5,6]. Radiology notes are free-text reports that document the results of various imaging studies, such as X-rays, CT scans, MRIs, and ultrasounds. While written in free text, these reports are often semi-structured and follow a consistent format depending on the imaging protocol. For instance, chest X-ray reports typically include sections like indication, comparison, findings, and impression [4,5,6]. Appendix B provides a list and description of all the features included in the final dataset.

Section 3.2: Cohort Building

ICD-10 code	**ICD-10 description**
	Alzheimer's Disease
G30.0	Early onset Alzheimer's Disease
G30.1	Late onset Alzheimer's Disease
G30.8	Other Alzheimer's Disease
G30.9	Alzheimer's Disease, unspecified
	Related dementias
F01	Vascular dementia without behavioral disturbance
F02	Dementia in other diseases classified elsewhere
F03	Unspecified dementia
G31.0	Frontotemporal dementia
G31.8	Other specified degenerative disease of nervous system
G31.9	Degenerative disease of nervous system, unspecified
R41.81	Age-related cognitive decline

Table 2: ICD-10 codes used to distinguish cases from controls

Case Cohort

The case cohort was curated from the MIMIC IV database by specifically selecting individuals diagnosed with AD using ICD-10 codes as defined in Table 2. Patients diagnosed with related dementia conditions were excluded in order to reduce heterogeneity in the case cohort. A total of 1,092 patients were identified as having an AD diagnosis and complete records described above. These patients formed the case cohort.

Control Cohort

The control cohort consisted of patients who did not have any documented AD or related dementias ICD-10 codes, as described in Table 2. These individuals served as unaffected comparisons to the AD case cohort. Propensity score matching using a logistic regression model was applied to identify 1,084 controls with complete records which were demographically similar to the AD cases defined above. A 1:1 matching approach was used to ensure each individual in the case cohort was paired with a control sharing similar baseline characteristics, based on the following key covariates: age, self-reported sex, self-reported race group, language, insurance type, and admission type. This method helps reduce confounding by balancing observed covariates across groups, thereby approximating the conditions of a randomized controlled trial within an observational dataset.

To assess the effectiveness of the 1:1 propensity score matching procedure, the distribution of key covariates across the treatment and control groups was visualized using overlaid histograms with kernel density estimates (KDEs). The distribution of the baseline characteristics for our cases and controls is shown in Table 3 and Appendix C.

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 Sex		
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

Table 3: Demographics information for study cohort

Section 3.4: Comorbidity Data

Structured indicators for key comorbidities were included in order to improve the predictive power of our model as some comorbid conditions may act as risk factors for cognitive decline, influencing the onset of AD. The comorbidities included were: stroke, myocardial infarction, peripheral vascular disease, cerebrovascular disease, diabetes mellitus, and cancer. These comorbidities were derived from ICD diagnosis data for patients in the matched case-control cohort. For each condition, we assigned a binary indicator (1 = condition present, 0 = not present) based on the presence of relevant ICD-9 or ICD-10 codes in the patient's medical history.

Section 4: Methodology

We developed machine learning models to predict whether an individual in the selected cohort had developed AD, using both their historical clinical notes and the demographic and comorbidity features defined earlier. In this study, we explored two modeling approaches: (1) summarizing concatenated notes using Gemini and incorporating the resulting summary as an additional feature in a traditional ML classification model, and (2) transforming structured tabular data into natural language sentences, appending them to the summarized notes, and feeding the combined text into Gemini for an LLM-based prediction. After generating and cleaning the prediction results, we evaluated modeling bias using commonly adopted fairness metrics. A comparative analysis of bias across methodologies is presented in Section 5.

Section 4.1: AD Prediction Methodology

Clinical Note Summarization

To generate structured, de-identified summaries of patient records for downstream modeling, we first concatenated all available clinical notes - including discharge summaries and radiology reports - into a single text field for each patient. However, patient medical history is frequently documented in clinical notes, such as discharge summaries, and the clinical notes available for many cases were documented after AD diagnosis. To mitigate information leakage into the AD prediction task, we designed prompts instructing the Gemini language model to explicitly mask

any mention of AD, its abbreviation, or medications commonly used to treat the condition. These prompts, along with the note text (truncated to 50,000 characters), were compiled into JSONL format and submitted to Gemini through Google Cloud's batch API with generation parameters set to a temperature of 0.4 and a maximum token output of 512. Upon receiving model responses, we extracted the generated summaries and aligned them with their corresponding patients using unique identifiers. The resulting summaries, now devoid of AD-related leakage, were merged into the main dataset as an additional feature for further analysis and prediction modeling.

The first sample prompt fed into Gemini is as follows:
"Please mask any mention of Alzheimer's Disease (including 'AD') and Related Dementias. Also, mask any medications used to directly treat these conditions. You are a trained neurologist. Your task is to provide a clear summary of the following note: {text}"

We also experimented with 3 different versions of the prompt, varying in technical instructions and output structure. Details on prompt 2 and 3 can be found in Appendix D.

Multi-modal ML prediction model

For the final prediction model, we combined categorical and textual features from the patient dataset, including demographic variables (e.g., age, self-reported sex, insurance group, language, and race), indicators for common comorbidities (such as diabetes, stroke, and cancer), and Gemini-generated clinical summaries. The textual data was encoded into numerical form using ClinicalBERT embeddings. To reduce dimensionality and mitigate overfitting, we applied Principal Component Analysis (PCA) to the combined feature set. A LightGBM classifier—selected for its efficiency and strong performance on high-dimensional, mixed-type data—was then trained using cross-validation across all three test prompts, with an additional PCA variation applied explicitly to Prompt 1. The resulting model outputs were used to generate final predictions for Alzheimer's disease and to evaluate overall model performance.

Gemini-only prediction model
Given the summarized notes explained above and a sentence formulation of the demographics and comorbidity tabular data, Gemini was prompted to act as a neurologist and provide a diagnosis of LIKELY_AD, POSSIBLE_AD, or UNLIKELY_AD and provide reasoning for its decision. A three way classification was used to discern Gemini's confidence in its predictions. In the prompt, LIKELY_AD was described as clear evidence consistent with AD, POSSIBLE_AD was described as the presence of some suggestive signs, but incomplete documentation, and UNLIKELY_AD was described as no indication of AD or dementia. The complete prompts are provided in Appendix E. Gemini predictions were very conservative and favored diagnosing POSSIBLE_AD over LIKELY_AD. Because of this, for evaluation of model performance and bias, we created a new prediction variable where POSSIBLE_AD and LIKELY_AD were categorized as 1 indicating a predicted AD case and UNLIKELY_AD as 0 indicating a predicted control. We evaluated the Gemini-only prediction model using three different predictions schemes: 1) cases predicted as only LIKELY_AD, 2) cases predicted as only POSSIBLE_AD, and 3) cases predicted as either LIKELY_AD or POSSIBLE_AD. In all schemes, only UNLIKELY_AD were used as controls.

Section 4.2: Statistical methodology for fairness evaluation
To rigorously evaluate model fairness, we focused on three key metrics: equal opportunity (measured via TPR), predictive parity (measured via precision), and equalized odds (measured via the joint parity of TPR and FPR). We applied bootstrapping to the sample data to estimate 95% confidence intervals for model performance, enabling intuitive visual comparisons. For equal opportunity and predictive parity, we conducted two types of statistical tests. At the subgroup-pair level, we used the two-proportion Z-test to identify significant differences between every pair of subgroups within a demographic attribute. At the overall group level, we applied the Chi-squared test to assess

whether the distribution of TPR or precision varied significantly across all subgroups of a given demographic feature. For equalized odds, we utilized the *equalized_odds_difference* function from the Fairlearn package, defined as the maximum absolute difference in TPR and FPR between any two subgroups. Thresholds for detecting bias were set at $p < 0.05$ for equal opportunity and predictive parity, and equalized odds difference (EOD) > 0.1, following commonly used industry guidelines.

Section 5: Results

Section 5.1 Multi-modal ML prediction model

We have the full results for all iterations of our models and prompts attached in Appendix F, and we discuss the training and testing set results for the fine-tuned LightGBM model using Prompt #1 to generate the Gemini summaries in greater detail below.

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

Table 4: Overall Model Performance on Training Set

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

Table 5: Overall Model Performance on Test Set

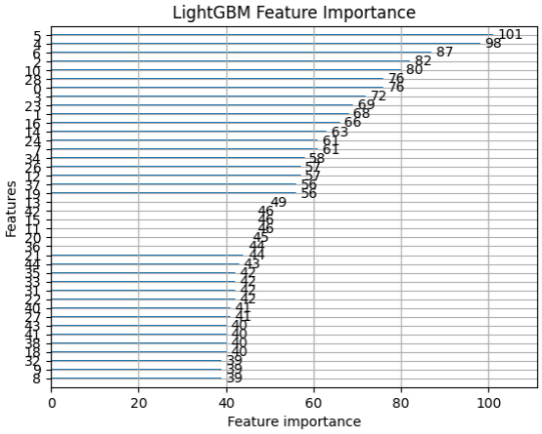


Figure 1: Top 40 Important Features in Multi-modal ML prediction model (Prompt 1, 80% PCA)

Despite fine-tuning parameters such as the learning rate, max depth, number of estimators, and others, the LightGBM model heavily overfits to the training data. There is almost a 20% drop in the TPR and precision metrics along with a roughly 21% increase in the FPR when moving from the training set to the testing set (Table 4,5). Although the model outperforms the benchmark that simply predicts all patients have AD - which achieves a TPR and FPR of 100% - the somewhat weak generalization performance could be attributed to the small sample size and dimensionality of the test set, with 56 features.

The LightGBM model uses all 10 categorical features, standardized age, and 45 of the top principal components to explain roughly 80% of variation in the textual summary data. Figure 1 shows the feature importance plot for this model. The plot indicates that many of the principal components ranked highly compared to tabular features in discerning whether an individual has AD. Consequently, the clinical note summaries play a key role in helping the model identify signals of cognitive decline or behavioral symptoms associated with AD.

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

Table 6: Fairness Metric by Race Subgroup (Train vs. Test)

We further performed subgroup analysis for different self-reported sex, age, and race groups to uncover potential model biases. Table 6 shows the race group fairness metric comparisons. Across racial subgroups, the model violates all three fairness criteria–equal opportunity, predictive parity, and equalized odds–which suggests signs of bias. The equal opportunity measure requires similar TPR between race groups, but there are wide variations in the test set results for the Hispanic/Latino group (0.143-0.714) and Asian group (0.500-1.000). Whereas, the White and African American groups have comparable ranges of 0.570-0.701 and 0.572-0.871, respectively. Thus, the model is less reliable at identifying true Alzheimer's patients in some racial subgroups compared to others, especially those with small sample sizes. Similarly, the predictive parity measure is also violated as we see greater uncertainty for the Hispanic/Latino and Asian groups relative to White, African American, and Other. Finally, the equalized odds metric is violated due to the lack of TPR and FPR parity. The Hispanic/Latino group has complete uncertainty with the FPR interval from 0.000 to 1.000 while the White, African American, and Other groups have narrower FPR ranges.

Subsequently, we conducted chi-square tests for each race group to assess whether the distribution of the TPR, FPR, or precision metrics differ significantly across subgroups. All the p-values were large, suggesting that we fail to reject our null hypothesis and conclude that there is not a statistically significant difference in the tested metrics between pairwise comparisons of all racial groups.

Even though the pairwise comparisons are not statistically significant, this does not directly indicate an absence of bias, especially given that we have small sample sizes and wide confidence intervals. In particular, as discussed above, our model displays disproportionate bias towards the Hispanic/Latino and Asian racial subgroups, which may present risks if applied in a clinical setting. This highlights important concerns regarding fairness and model reliability.

Section 5.2: Gemini-only prediction model

We first examined model performance by combining cases predicted as either *LIKELY_AD* or *POSSIBLE_AD* into a single positive class. As the train-test split was no longer preserved in the final dataset, model evaluation was conducted on the full original cohort of 2,176 patients. Table 7 and 8 presents the overall model performance metrics, followed by a summary of fairness violations identified across major demographic groups.

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

Table 7: Overall Model Performance on Gemini-only Prediction

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

Table 8: Summary of Fairness Metrics by Demographic Group (X represents fairness violation)

Equal Opportunity (TPR Parity):
A chi-squared test across the four age groups (<70, 70–80, 80–90, >90) revealed a statistically significant difference in TPR distribution ($p = 0.0050$), indicating group-level disparities in the model's ability to correctly identify positive cases across age strata. Subsequent pairwise comparisons confirmed this finding, with particularly large gaps in sensitivity between individuals under age 70 and those over 90 (TPR = 0.485 vs. 0.690, $p = 0.0004$), as well as between <70 and 80–90 (TPR = 0.485 vs. 0.640, $p = 0.0036$), and between 70–80 and <70 ($p = 0.0117$).

For other demographic variables - including self-reported sex, insurance group, and language—no statistically significant differences in TPR were observed, suggesting no broad disparities in sensitivity within these groups. However, in the race-related subgroups, while the overall group-level chi-squared test did not indicate significant bias, pairwise testing revealed a notable disparity between African American and Hispanic/Latino individuals (TPR = 0.681 vs. 0.510, $p = 0.0328$), suggesting that sensitivity imbalances may still exist within specific subgroup comparisons even when not detected at the broader group level.

Predictive Parity (Precision Parity):
Age groups also showed significant disparities in predictive parity. A chi-squared test across the four age bins revealed a statistically significant difference in precision ($p = 0.0096$). Pairwise comparisons indicated that individuals aged 70–80 had higher precision than those over 90 (0.788 vs. 0.658, $p = 0.0023$), and individuals under 70 also outperformed the >90 group (0.803 vs. 0.658, $p = 0.0292$), suggesting reduced predictive reliability in the oldest age bracket. No significant precision differences were observed across self-reported sex groups. However, in the insurance category, a significant disparity was detected between individuals on Medicare and those classified under "Other" insurance types ($p = 0.0259$), indicating uneven performance based on insurance coverage.

Equalized Odds:
When jointly considering TPR and FPR, several demographic attributes—particularly self-reported sex and race - exhibited substantial disparities. In multiple categories, including age, insurance group, and racial subgroups, the lower bound of the 95% confidence interval for the EOD exceeded the threshold of 0.1, indicating systematic inequities in error distribution across groups. Consistent with the patterns observed in equal opportunity and predictive parity analyses, the most pronounced bias was found between individuals under age 70 and those over 90. Among this subgroup pair ($n = 691$), the bootstrapped EOD ranged from 0.163 to 0.326, underscoring a considerable imbalance in the model's treatment of extreme age groups.

Furthermore, because we asked Gemini to provide differing levels of confidence with LIKELY_AD and POSSIBLE_AD, with LIKELY_AD being higher confidence, we were able to evaluate fairness metrics stratified by Gemini's confidence in its prediction. We found that the majority of the biases found in the Gemini model were driven by low-confidence predictions (i.e., POSSIBLE_AD predictions) (Appendix G). This suggests that limiting predictions to high-confidence ones may be a potential bias mitigation strategy.

Overall, the analysis revealed statistically significant fairness violations across all three metrics, with evidence of bias both between subgroup pairs and across entire demographic categories. These results highlight critical equity concerns that warrant mitigation in future modeling efforts.

Section 5.3: Model Comparison and Discussions

All models achieved classification accuracy above the 50% benchmark, confirming their predictive utility; however, they consistently exhibited bias in equalized odds and, in some cases, in the other two fairness metrics as well. Besides, the choice of summarization prompt and model architecture had a notable impact on fairness outcomes.

Prompt 1, a simplified instruction used within a traditional machine learning pipeline, achieved better-than-baseline accuracy and exhibited no significant bias in equal opportunity or predictive parity. However, equalized odds analysis revealed age-related disparities, particularly affecting the youngest and oldest patients. This suggests that while simplification may reduce leakage from explicitly masked terms, it may not fully capture clinically relevant subtleties across age groups.

In contrast, Prompt 3, a fine-tuned and clinically structured instruction set, led to the best overall test set performance as well as fairness profile. This model effectively mitigated subgroup disparities, particularly eliminating age-group bias in equalized odds, which was prominent in other configurations. It achieved a balanced performance (TPR = 0.652, FPR = 0.292, Precision = 0.708) and preserved critical clinical information (e.g., cognitive symptoms and functional status) without overexposing sensitive AD-related terms. This structured prompt likely facilitated more equitable representation across demographic groups by encouraging consistent, medically grounded summaries.

Experiments with dimensionality reduction revealed a potential trade-off between performance and fairness. Increasing PCA components from covering 50% to 80% variance improved model performance on both training and testing dataset but introduced potential overfitting in model and new equalized odds bias in language groups, highlighting the sensitivity of fairness to latent feature complexity.

The Gemini-only model, which bypassed structured downstream modeling and relied solely on LLM-generated outputs for final prediction, achieved performance above the 50% threshold but exhibited pronounced bias—particularly with respect to age. One possible explanation is that Gemini was trained on broad, real-world data, which likely introduced pre-existing societal biases. Additionally, lacking train-test calibration and access to model reasoning for prompt-level tuning limited fairness optimization and our ability to mitigate bias effectively.

These findings demonstrate that among all 3 fairness metrics, equalized odds is the hardest to achieve. They also demonstrate that traditional ML models, when paired with well-engineered prompts and structured evaluation, offer greater potential to balance both predictive accuracy and demographic fairness compared to LLM-only pipelines.

Section 6: Discussion

Data Limitations

To improve the generalizability and robustness of our findings, future work will focus on addressing several data-related limitations. Our current dataset consists of approximately 2,000 ICU patients from Beth Israel Deaconess Medical Center, which presents constraints in both sample size and demographic diversity. We plan to increase the dataset size and broaden the patient population by incorporating data from additional hospitals and care settings, particularly non-ICU environments. Furthermore, our analysis currently draws only from discharge summaries and radiology notes, which may not fully reflect the dynamic clinical context. Future efforts will involve integrating nursing notes, which are recorded more frequently and may provide more granular, real-time insights to enhance model performance and fairness assessments.

Model Improvement

To enhance model reliability, future work will explore strategies to reduce overfitting in the LightGBM model, including more robust regularization techniques and cross-validation frameworks. Future work could involve tracing

important PCA components back to their embedding origins to uncover which semantic elements in the clinical notes most influence prediction outcomes. Additionally, we observed that the Gemini-based model failed to generate outputs for some patients, likely due to safety filters inherent in its architecture. This motivates future exploration into developing or adapting large language models specifically for healthcare applications - models that maintain safety while being capable of consistently generating summaries and predictions from sensitive clinical data.

Fairness Evaluation and Mitigation

Future work will broaden the scope of fairness evaluation beyond the current use of equal opportunity, predictive parity, and equalized odds. We intend to incorporate additional metrics such as demographic parity, calibration within groups, and treatment equality to provide a more comprehensive and nuanced understanding of model behavior across diverse subpopulations. While initial results indicate minimal bias - primarily under equalized odds - this may vary depending on the final model selected, particularly if optimizing for performance introduces unintended disparities.

To better understand and address the origins of bias, other than continuous prompt engineering, one promising direction is to examine the internal decision-making processes of the Gemini model, evaluating how sensitive attributes may influence its outputs. In parallel, for the multi-modal approach, we plan to investigate bias mitigation strategies across the entire machine learning pipeline, including:

- Pre-processing (e.g., reweighting, resampling, and feature transformation),
- In-processing (e.g., adversarial debiasing [9] and fairness-aware loss functions), and
- Post-processing (e.g., threshold optimization, and reject option classification [10]).

These initiatives will be supported by continuous fairness auditing, ensuring that equity is maintained as models are updated with new data and deployed in evolving real-world contexts.

Section 7: Acknowledgments & Member Contributions

We would like to thank our clinical mentors—Yingnan He, Yu Leng, and Dr. Sudeshna Das—as well as our professor, Dr. Peter Szolovits, without whom this study would not have been possible.

Anya merged the MIMIC-IV and MIMIC-IV-NOTE datasets to create a comprehensive case-control cohort, identified patients with complete data, and developed the Gemini-based AD prediction model through engineered prompts for summarization and diagnosis. She also analyzed how Gemini's prediction confidence affected fairness outcomes. Sierra led the cohort selection logic, performed 1:1 case-control matching, and evaluated model performance for both the multimodal and Gemini-only models across all fairness metrics. She also trained multiple model variants, conducted statistical bias testing, and guided model comparison and interpretation. Nidhish integrated comorbidity data via ICD codes, built the Gemini summarization pipeline, and fine-tuned prompt instructions. He also generated and interpreted feature importance plots and analyzed bias sources using fairness metrics and statistical tests. Priya conducted a literature review on AD detection models, led exploratory data analysis on MIMIC-IV, and contributed to Gemini prompt engineering. She evaluated how prompt variation impacted clinical note summaries and assessed its effect on downstream LightGBM model performance.

Section 8: Code Availability

All code used in the report is available at https://github.com/xgong75/Characterizing-Bias-in-AD-Detection.

References

1. Scheltens, Philip et al. Alzheimer's disease. *Lancet.* **397**,10284 (2021). https://doi:10.1016/S0140-6736(20)32205-4
2. Chen, IY., et al. Ethical Machine learning in healthcare. *Annu. Rev. Biomed. Data. Sci.* **4**, 123-144 (2021). https://doi.org/10.1146/annurev-biodatasci-092820-114757
3. Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x
4. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.
5. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (version 3.1). *PhysioNet*. https://doi.org/10.13026/kpb9-mt58.
6. Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). *PhysioNet*. https://doi.org/10.13026/1n74-ne17.
7. Shankar, R., Bundele, A. & Mukhopadhyay, A. Natural language processing of electronic health records for early detection of cognitive decline: a systematic review. *npj Digit. Med.* **8**, 133 (2025). https://doi.org/10.1038/s41746-025-01527-z
8. Leng, Y., He, Y., et al. Evaluating GPT's capability in identifying stages of cognitive impairment from electronic health data. *arXiv*. (2025). https://doi.org/10.48550/arXiv.2502.09715
9. Yang, J., Soltan, A.A., et al. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. npj Digital Medicine. (2023). https://www.nature.com/articles/s41746-023-00805-y
10. Goethals, S., Calders, T., et al. Beyond accuracy-fairness: Stop evaluating bias mitigation methods solely on between-group metrics. arXiv. (2024). https://arxiv.org/html/2401.13391v1

APPENDIX

Appendix A: Derived Features

1) Age

The patient's age at the time of admission was calculated using 'anchor_year', 'anchor_age, and 'admit_year'. The 'anchor_year' represents a de-identified year tied to a known age, 'anchor_age', and 'admit_year' represents a de-identified year tied to when the patient was admitted. The patient's age was calculated as follows:
age = admityear - anchor_year + anchor_age

2) Race

To reduce sparsity and ensure consistent categorization across the dataset, detailed self-reported race and ethnicity entries were consolidated into broader categories through a custom mapping race_map1 as described below.

- Subgroups beginning with "White" (e.g. "White - Russian", "White - Other European", etc.) were grouped under the label "White".
- Subgroups beginning with "Hispanic/Latino" (e.g., "Hispanic/Latino - Mexican", "Hispanic/Latino - Dominican", etc.) were grouped in a unified "Hispanic/Latino" category.
- Subgroups beginning with "Asian" (e.g., "Asian - Chinese", "Asian - SouthEast Asian", etc.) were combined under the label "Asian".
- Ambiguous responses (e.g., "Unknown," "Patient Declined to Answer", "Multiple Race/Ethnicity") were grouped into the "Other" category.
- All less frequent self-reported race categories in the dataset with less than 10 total patients (e.g. "Cape Verdean" (n=2), "Caribbean Island" (n=1), etc.) were also merged into the "Other" category.

Only the four major race categories ("White", "African American", "Asian", and "Hispanic/Latino") were maintained as distinct groups.

3) Language

If a patient's language information was missing, it was classified as "Unknown". The patient's language was then categorized as "English" or "Non-English" under 'language_group'.

4) Insurance

Similarly, the patient's insurance information was categorized as Medicare, Medicaid, or Other, with "Private," "Other," and missing values grouped under "Other," in the derived feature labeled insurance_group.

Appendix B: Final Dataset Features

Feature	Description	Type
Subject ID	an integer number identifying a particular patient	structured, raw
Sex	patient self-reported sex	structured, raw
Age	patient age calculated using $\text{anchor}_{a}geand\text{anchor}_{y}ear$	structured, derived
Admission Type	label classifying the type of admission that may be used as proxy for urgency	structured, raw
Insurance Group	patient's health insurance plan	structured, derived
Language	patient's self-reported language (English or non-English)	structured, derived
Race Group	mapped detailed self-reported race categories into broader standardized groups	structured, derived
Stroke History	binary indicator for documented history of stroke	structured, derived
Myocardial Infarction	binary indicator for documented history of myocardial infarction	structured, derived
Peripheral Vascular Disease	binary indicator for documented history of peripheral vascular disease	structured, derived
Cerebrovascular Disease	binary indicator for documented history of cerebrovascular disease	structured, derived
Diabetes Mellitus	binary indicator for documented history of diabetes mellitus	structured, derived
Cancer	binary indicator for documented history of cancer	structured, derived
Clinical Notes	free-text content of discharge and radiology notes combined and summarized	unstructured, derived (combined)

Appendix Table 1: Overview of Final Dataset Features

Appendix C: Additional Study Population Characteristics

Category	# Cases	# Controls
Admission Type		
Ambulatory Observation	4	8
Direct Emergency	24	28
Direct Observation	15	25
Elective	11	20
EU Observation	90	87
EW Emergency	296	409
Observation Admit	534	354
Surgical Same-Day Admit	30	56
Urgent	88	97

Appendix Table 2: Distribution of admission type among AD cases and controls

Appendix D: Prompt Engineering for Multi-modal ML Prediction Model

We further explored the impact of prompt engineering on model outputs by designing two more detailed prompts. These prompts - and their corresponding results - are detailed below.

Prompt 2

Prompt 2 was structured as follows:

prompt = f"""You are a clinical documentation specialist assisting in a medical research project aiming to predict Alzheimer's Disease. Please carefully review the following patient notes (which may consist of multiple entries combined). Perform the following steps:

*1. **Mask** (i.e., replace with [MASKED]) any explicit diagnosis or direct mention of:*

- *Alzheimer's Disease (AD)*
- *Dementia (only if directly diagnosed; general symptoms should be preserved)*
- *Medications specifically prescribed for Alzheimer's Disease (e.g., Donepezil, Memantine, Rivastigmine, Galantamine).*

*2. **Do not** mask or remove:*

- *Symptoms that could indicate cognitive decline, memory loss, confusion, disorientation, difficulty with language, executive dysfunction, or behavior changes.*
- *General clinical observations that might hint at early-stage Alzheimer's or other neurological impairments.*

*3. **Summarize** the patient's clinical history **as a unified record** - do not separate summaries by note.*

- *Focus on accurately describing cognitive, neurological, psychiatric, and functional status.*
- *Include relevant medical history, symptoms, and treatments (except masked medications).*
- *Use clinical language appropriate for a medical professional audience.*

Here are the patient's combined notes: {text}"""

Prompt 3

Prompt 3 was structured in order to remove the term "[MASKED]" from the summaries generated from Prompt 2 in order to ensure that the prediction model does not inadvertently learn from or rely on this placeholder during training.

prompt = f"""You are a clinical documentation specialist assisting in a medical research project aiming to predict Alzheimer's Disease. Please carefully review the following patient notes (which may consist of multiple entries combined). Perform the following steps:

*1. **Mask** any explicit diagnosis or direct mention of:*

- *Alzheimer's Disease (AD)*
- *Dementia (only if directly diagnosed; general symptoms should be preserved)*
- *Medications specifically prescribed for Alzheimer's Disease (e.g., Donepezil, Memantine, Rivastigmine, Galantamine).*

*2. **Do not** mask or remove:*

- *Symptoms that could indicate cognitive decline, memory loss, confusion, disorientation, difficulty with language, executive dysfunction, or behavior changes.*
- *General clinical observations that might hint at early-stage Alzheimer's or other neurological impairments.*

3. Summarize the patient's clinical history as a unified record, not by individual notes.

Follow this structure for every patient to ensure consistency:

- *Cognitive Status: Describe memory, attention, language, and executive function findings.*
- *Neurological Findings: Include motor/sensory exam, reflexes, gait, coordination, and imaging results if mentioned.*
- *Psychiatric Symptoms: Summarize mood, behavior, and psychiatric diagnoses or concerns.*
- *Functional Abilities: Describe the patient's level of independence in ADLs/IADLs and any changes noted.*
- *Relevant Medical History: Include comorbidities and chronic conditions that may affect cognitive or functional status.*
- *Symptoms & Treatments: Describe current symptoms and any treatments received (excluding masked medications).*

Use precise clinical language suitable for a medical professional audience.

Here are the patient's combined notes: {text}"""

Appendix E: Gemini-only prediction model prompts

Prompt for masking AD in clinical notes

"*You are a clinical documentation specialist assisting in a medical research project aiming to predict Alzheimer's Disease from clinical notes. Please carefully review the following clinical notes and remove phrases with explicit diagnosis or direct mention Alzheimer's Disease (AD), dementia (only if directly diagnosed, general symptoms should be preserved), and medications specifically prescribed for Alzheimer's Disease (e.g., Donepezil, Memantine, Rivastigmine, Galantamine).*
A phrase is defined as any series of words contained within two commas (,), periods (.), parentheses (() or []), or single line (indicated by \n).
Do not remove symptoms that could indicate cognitive decline, memory loss, confusion, disorientation, difficulty with language, executive dysfunction, or behavior changes.
Do not remove general clinical observations that might hing at early-stage Alzheimer's or other neurological impairments.
Clinical notes: {text}"

Prompt for predicting AD

"You are a neurologist tasked with diagnosing Alzheimer's disease based on a summary of a patient's clinical notes. Please read the following clinical notes summaries and assess whether there is documented evidence suggestive of Alzheimer's Disease (AD). Consider cognitive symptoms such as progressive memory loss, disorientation, language difficulties, and behavioral changes as well as test results such as cognitive assessments or neuroimaging findings indicative of brain atrophy or AD markers. Do not include temporary confusion from infection, medication, or unrelated acute illness.
Respond with one of the following labels:
- LIKELY_AD: clear evidence consistent with Alzheimer's Disease
- POSSIBLE_AD: some suggestive signs, but incomplete documentation
- UNLIKELY_AD: No indication of Alzheimer's Disease or dementia
Also provide a 1-2 sentence justification summarizing the key evidence from the note.
The following is the clinical notes summary for a {age} year old {race} {sex} who is {marital_status}.
They were admitted to the hospital as {admission}, have {insurance} insurance, and are {language} speaking.
They have the following disease history: {disease}
Clinical Notes: {text}"

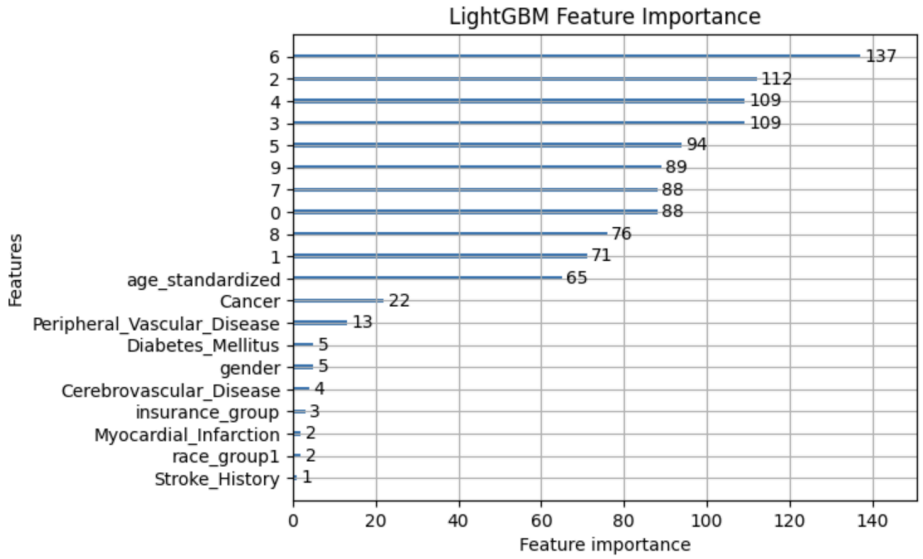
Appendix F: Results of using embeddings of summaries from prompt 2 and 3

Prompt 1 Results (50% Variance for PCA)

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

Race Group	Equal Opportunity Range	FPR Range	Predictive Parity Range	Sample Size
		Training Set		
White	0.736–0.799	0.153–0.211	0.771–0.829	1267
African American	0.659–0.828	0.101–0.235	0.768–0.892	210
Other	0.716–0.910	0.171–0.381	0.690–0.865	136
Hispanic/Latino	0.666–0.907	0.074–0.282	0.703–0.920	75
Asian	0.833–1.000	0.018–0.223	0.796–0.987	50
		Testing Set		
White	0.516–0.662	0.305–0.438	0.569–0.715	327
African American	0.439–0.830	0.057–0.356	0.512–0.938	54
Hispanic/Latino	0.348–0.947	0.223–1.000	0.408–0.900	15
Other	0.389–0.913	0.071–0.537	0.375–0.923	27
Asian	0.308–1.000	0.000–0.667	0.241–1.000	13

Appendix Table 3: Summary of Fairness Metric Violations by Demographic Group with Race Snapshot



Appendix Figure 1: Feature importance for multi-modal model with prompt 1 embeddings

21 PCA features

Prompt 2 Results (80% Variance for PCA)

Metric	Value
True Positive Rate (TPR)	0.802
False Positive Rate (FPR)	0.158
Precision	0.834
Sample Size	1740

Appendix Table 4: Overall model performance on training set using prompt 2 embeddings

Metric	Value
True Positive Rate (TPR)	0.586
False Positive Rate (FPR)	0.306
Precision	0.675
Sample Size	436

Appendix Table 5: Overall model performance on test set using prompt 2 embeddings

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

Appendix Table 6: Summary of Fairness Metric Violations by Demographic Group

Prompt 3 Results (80% Variance for PCA)

Metric	Value
True Positive Rate (TPR)	0.945
False Positive Rate (FPR)	0.0594
Precision	0.940
Sample Size	1740

Appendix Table 7: Overall model performance on training set using prompt 3 embeddings

Metric	Value
True Positive Rate (TPR)	0.652
False Positive Rate (FPR)	0.292
Precision	0.708
Sample Size	436

Appendix Table 8: Overall model performance on test set using prompt 3 embeddings

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

Appendix Table 9: Summary of Fairness Metric Violations by Demographic Group (X represents fairness violation)

Of the three prompts, Prompt 3 delivers the best performance on the test set, while Prompt 2 lags behind across all key metrics. Prompt 3 introduces detailed, structured formatting for summarization, resulting in more consistent outputs than Prompt 2. This suggests that greater consistency in summary generation may enhance model performance.

However, the improvement achieved by Prompt 3 over Prompt 1 is relatively modest, especially considering the simplicity of Prompt 1. This suggests that increasing prompt complexity does not always translate to substantially better model performance.

Appendix G: Results of Gemini-based model stratified by confidence

Metric	Value
cases = LIKELY_AD only	
True Positive Rate (TPR)	0.467
False Positive Rate (FPR)	0.091
Precision	0.811
Accuracy	0.709
Sample Size	1656
cases = POSSIBLE_AD only	
True Positive Rate (TPR)	0.460
False Positive Rate (FPR)	0.179
Precision	0.656
Accuracy	0.668
Sample Size	1743

Appendix Table 10: Overall Model Performance on Gemini-only Prediction

Demographic Group	Equal Opportunity	Predictive Parity	Equalized Odds
cases = LIKELY_AD only			
Age Group	✓	✓	✓
Sex	✓	✓	✓
Insurance Group	✓	✓	✓
Language Group	✓	✓	✓
Race Group	✓	✓	✗
cases = POSSIBLE_AD only			
Age Group	✗	✗	✗
Sex	✓	✓	✓
Insurance Group	✓	✓	✗
Language Group	✓	✓	✓
Race Group	✓	✓	✗

Appendix Table 11: Summary of Fairness Metric Violations by Demographic Group (X represents fairness violation)