NeuroScope: Feature Analysis for Alzheimer's Risk

Devika Band **Ethan Ngo** **Raja Muhammad Shayan** **Freya Shah**

CSE 575: Statistical Machine Learning
Arizona State University

1 Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that currently lacks a cure, making early detection essential for improving patient outcomes through timely interventions and care planning. As diagnosis is often delayed until significant cognitive decline has already occurred, there is a pressing need for data-driven approaches that can support early identification using accessible clinical and behavioral features [13, 2]. Our project aims to bridge this gap by developing an interpretable machine learning pipeline that predicts Alzheimer's risk and identifies the most influential predictive features using real-world clinical datasets.

We, Group NeuroScope, focused on designing a solution that prioritizes both predictive performance and explainability, ensuring it can be trusted and adopted by clinical practitioners. Our objective was to assess and compare multiple classification models—Logistic Regression [10], Random Forest [11], and XGBoost [5]—on a curated dataset of over 800 patients aged 60 to 90 [12]. We placed particular emphasis on optimizing recall for Alzheimer's detection, minimizing false negatives that could delay necessary treatment.

To ensure clinical relevance and transparency, we incorporated explainable AI techniques such as SHAP (SHapley Additive exPlanations) [16] and LIME (Local Interpretable Model-Agnostic Explanations) [18] to interpret model predictions at both global and local levels. Additionally, we conducted fairness evaluations across gender, age, and ethnicity subgroups to ensure responsible deployment across diverse populations.

As a team:
Devika led data cleaning, feature engineering, and exploratory analysis.
Ethan trained and evaluated the Random Forest model and contributed to visual diagnostics.
Freya fine-tuned and interpreted the XGBoost model with SHAP and LIME explanations.
Shayan built and evaluated the Logistic Regression baseline, conducted fairness analysis, and led performance comparisons across models and datasets.

Together, we developed a reproducible and interpretable ML pipeline that not only achieves strong classification performance but also provides insights into the most important risk factors for Alzheimer's. We hope this work can inform future development of clinical decision-support tools and contribute to more equitable and proactive healthcare delivery.

2 Methods

2.1 Dataset Description

We used a real-world, structured dataset containing clinical records of over 800 patients aged between 60 and 90 [12]. Each record includes more than 30 features encompassing demographics (e.g., Age, Gender, Ethnicity), lifestyle habits (e.g., Smoking, Physical Activity, Sleep Quality), medical history (e.g., Hypertension, Cardiovascular Disease, Depression), and cognitive/behavioral assessments (e.g., MMSE, Functional Assessment, ADL, Memory Complaints, Confusion). The binary target variable indicates Alzheimer's diagnosis: AD (1) or No AD (0). After verifying the absence of missing values

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

and constant columns, we partitioned the data into 60% training, 20% validation [15], and 20% testing splits, maintaining class distribution through stratification.

2.2 Feature Processing

We categorized features as continuous or categorical and identified potential class imbalance in several binary features. Numerical features were standardized using z-score normalization (e.g., Age, MMSE), while categorical features such as Gender and Smoking were encoded using binary indicators [8].

Based on domain knowledge, correlation heatmaps, and exploratory visualizations, we selected 16 core predictors for model training. These included cognitive/functional scores (e.g., MMSE, ADL, FunctionalAssessment), behavioral symptoms (e.g., Confusion, Forgetfulness, Personality Changes), and medical risk indicators (e.g., Depression, Head Injury, Diabetes). Irrelevant metadata such as PatientID and DoctorInCharge were excluded.

The pairwise Pearson correlation coefficients between each of the dataset's numerical features are displayed in this heatmap. Lower cognitive and functional scores are strongly linked to Alzheimer's disease cases, as evidenced by the substantial negative correlations found between the Alzheimer's diagnosis label and the Mini-Mental State Examination (MMSE), Activities of Daily Living (ADL), and Functional Assessment scores.

Forgetfulness, confusion, and memory complaints all had positive correlations with the diagnosis and with each other, indicating a cluster of behavioral symptoms common to the progression of dementia.

The majority of clinical history characteristics, such as diabetes, cardiovascular disease, and hypertension, show weak to moderate associations, indicating that their effects may be more indirect or non-linear.

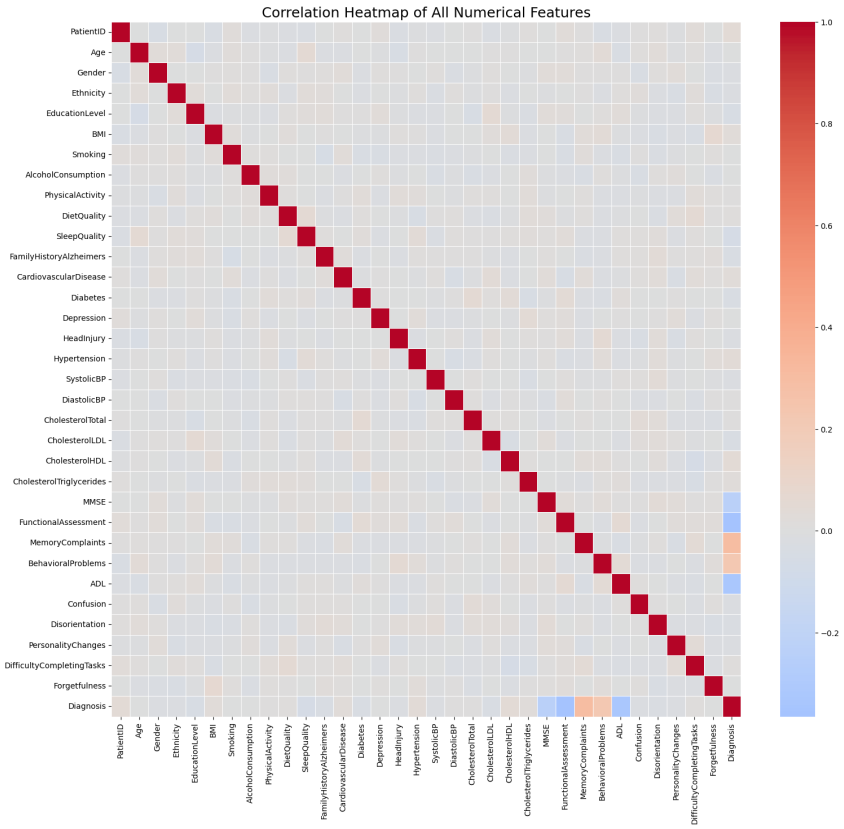


Figure 1: Correlation heatmap of all numerical features.

2.3 Modeling Pipeline

We trained three classifiers: Logistic Regression, Random Forest, and XGBoost.

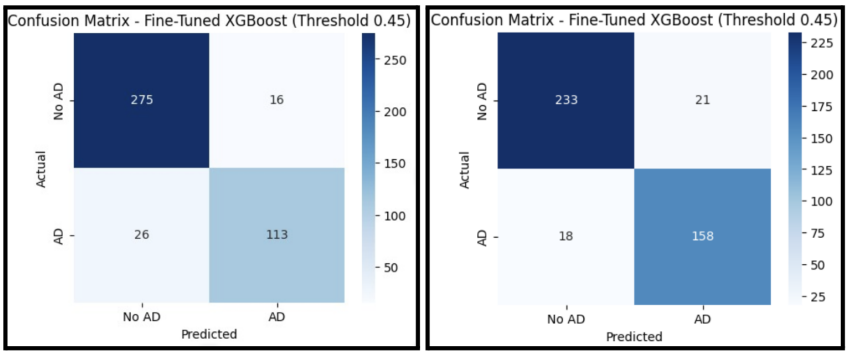
Logistic Regression (LR): We used scikit-learn's `LogisticRegression` with the `liblinear` solver. Z-score normalization was applied to all continuous features. To increase sensitivity for Alzheimer's prediction, we lowered the classification threshold from 0.5 to 0.3 after validation-based tuning.

Random Forest (RF): Implemented via scikit-learn's `RandomForestClassifier`, this ensemble method averaged predictions from multiple decision trees. We tuned hyperparameters including tree depth and number of estimators. The classification threshold was also set at 0.3 to increase recall.

XGBoost (XGB): As our final and best-performing model, we implemented XGBoost using `XGBClassifier`. We performed 5-fold stratified cross-validation on the training data and used `scale_pos_weight = 1.2` to handle class imbalance. After iterative tuning, we set the decision threshold to 0.45 to maximize recall while limiting false positives. Key hyperparameters included:

- `max_depth`: 4–6
- `learning_rate`: 0.05
- `n_estimators`: 100–150
- `subsample`: 0.8
- `colsample_bytree`: 0.7

The performance of a refined XGBoost model with a threshold of 0.45 is displayed in the two confusion matrices, one for the test data and one for the validation data. Strong performance is shown by both matrices, which show high true positives (113 for the test, 158 for validation) and true negatives (275 for the test, 233 for validation), suggesting that both AD and No AD cases are well classified. Although they are modest, the validation set's false positive and false negative rates are marginally higher, indicating some overfitting. All things considered, the model exhibits good generalization and a great capacity to forecast both AD and No AD in unknown data.



Test dataset Validation dataset

Figure 2: Comparison of XGBoost Confusion Matrix on Test vs. Validation Data.

2.4 Explainability: SHAP and LIME

To ensure interpretability of our final XGBoost model:

SHAP (SHapley Additive exPlanations): We generated global SHAP summary plots to rank the most influential features [9, 7]. SHAP values helped quantify each feature's marginal contribution to the model's prediction, revealing that FunctionalAssessment, ADL, MMSE, and MemoryComplaints were the most predictive. Higher feature values typically increase the likelihood of AD, as seen by the color gradient, which goes from blue (low values) to red (high values), which shows how feature values affect the model's output. The most important elements influencing the model's predictions can be found using this figure.

LIME (Local Interpretable Model-Agnostic Explanations): We also used LIME to analyze individual patient predictions [17]. These local visualizations showed how specific feature values influenced the prediction outcome (e.g., AD vs. No AD) for a given sample, further aiding clinical transparency. From the given figure above, we can see that the model is predicting the likelihood of Alzheimer's disease with a 16% chance of having no AD and an 84% chance of having AD. ADL is the most significant aspect; a score of ≤ 2.57, with a value of 2.00, suggests a higher risk of AD. The prediction is significantly influenced by behavioral problems and MMSE (Mini-Mental State Examination) scores, among other important aspects. The approach emphasizes that the main markers of Alzheimer's disease are cognitive symptoms and functional deficits, whereas physical health issues like hypertension are less significant.

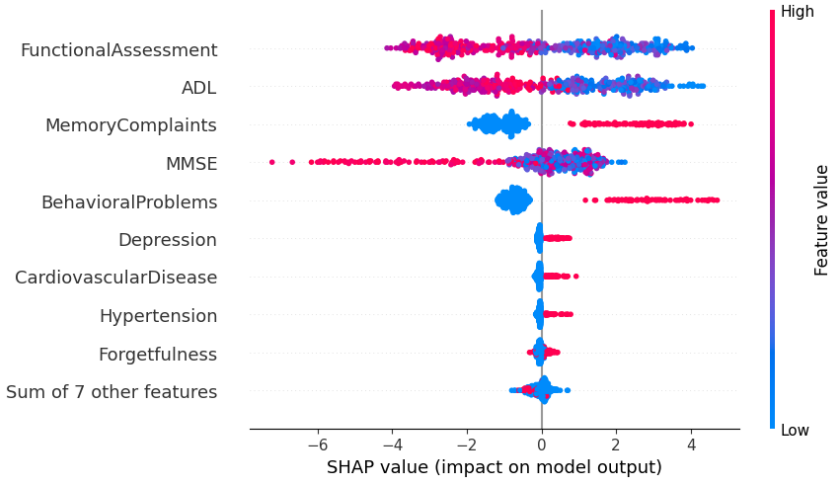


Figure 3: XGBoost Val Data SHAP Plot.

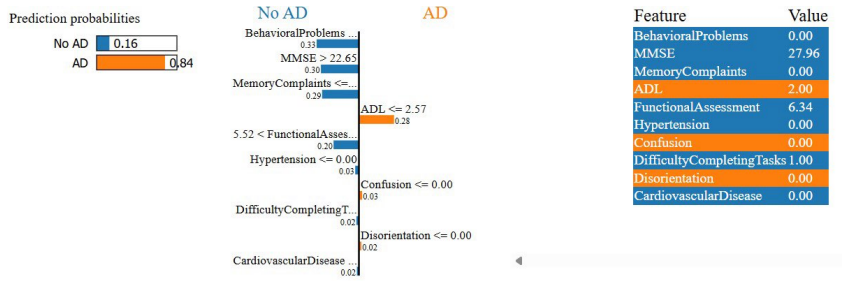


Figure 4: XGBoost Val Data LIME Plot.

2.5 Fairness Evaluation

We conducted a fairness audit on both validation and test splits [1] by stratifying model performance across three dimensions: Gender, Age Group, and Ethnicity. For each subgroup, we computed accuracy, precision, recall, and F1-score. This helped reveal any systematic performance disparities, informing discussions around responsible AI deployment in healthcare.

From the given bar chart comparison above we can see that both male and female F1-scores and recall are balanced, indicating that the model is gender-neutral.

We assessed equity by age, gender, and ethnicity. While results based on age and ethnicity showed greater discrepancy, especially underperformance for older patients and those in underrepresented ethnic groups, female patients showed somewhat superior predictive performance on the validation split.

Gender differences considerably decreased and overall age-group performance increased when the test split was administered again. Disparities remained, nevertheless, for Asian and African American groupings, suggesting the necessity of continuous fairness auditing.

To put it briefly, our approach works consistently overall, but if it is included into actual healthcare systems, precautions will be needed to guarantee fair results.

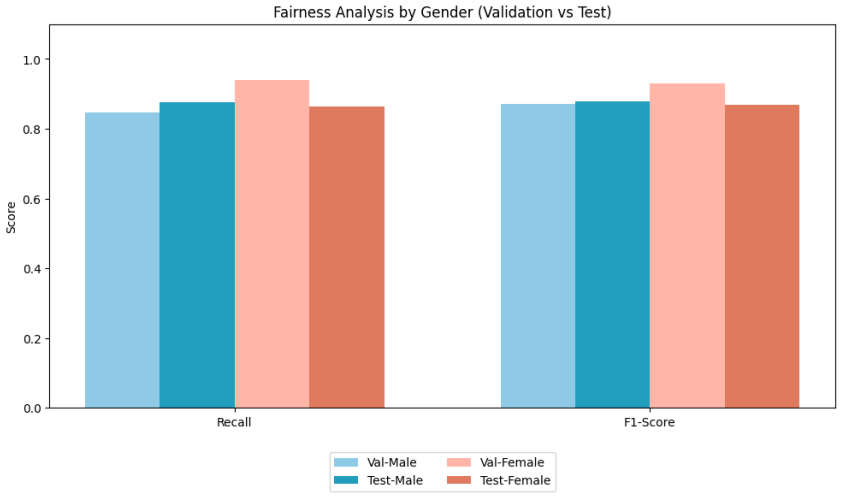


Figure 5: Fairness Analysis by Gender (Validation vs. Test).

3 Experiments and Results

3.1 Experimental Setup

All models were implemented in Python using the scikit-learn and XGBoost libraries. Experiments were conducted on a local machine with 16GB RAM. We used a stratified 60/20/20 split for training, validation, and testing to preserve class proportions [15]. Each model was evaluated using accuracy, precision, recall, F1-score, and AUC [14]. Additionally, we tuned decision thresholds to prioritize recall, which is critical for high-stakes medical diagnostics.

3.2 Baseline: Logistic Regression

Logistic Regression served as our interpretable baseline [10]. Initial results using a 0.5 threshold underperformed on Alzheimer's (AD) recall, leading us to reduce the threshold to 0.3. This adjustment improved AD recall from 64% to 80%, while maintaining 78% accuracy and 0.71 precision. The ROC-AUC score of 0.87 indicated strong separability between AD and No AD classes as shown in Figure 6 below. A confusion matrix revealed 36 false negatives, underscoring the challenge of sensitivity in clinical prediction tasks.

The confusion matrix also showed 140 true positives and 196 true negatives, but concerningly, it included 38 false negatives—cases where individuals with Alzheimer's were missed by the model. Given the clinical stakes, this is a substantial number, as undetected cases could delay intervention during a crucial treatment window. While the model accepted 58 false positives, the lowered threshold aimed to prioritize sensitivity. This trade-off reflects our project's core objective: minimizing missed diagnoses, even at the cost of slightly reduced specificity—an approach justified by the real-world imperative of early detection in high-risk medical conditions.

3.3 Random Forest

Random Forest offered a robust, nonlinear baseline [11]. After tuning max depth and the number of estimators, the model achieved 90.23% accuracy, with an F1-score of 0.87 for the AD class. However, despite high precision (0.97), recall remained modest (0.78) due to a conservative prediction tendency.

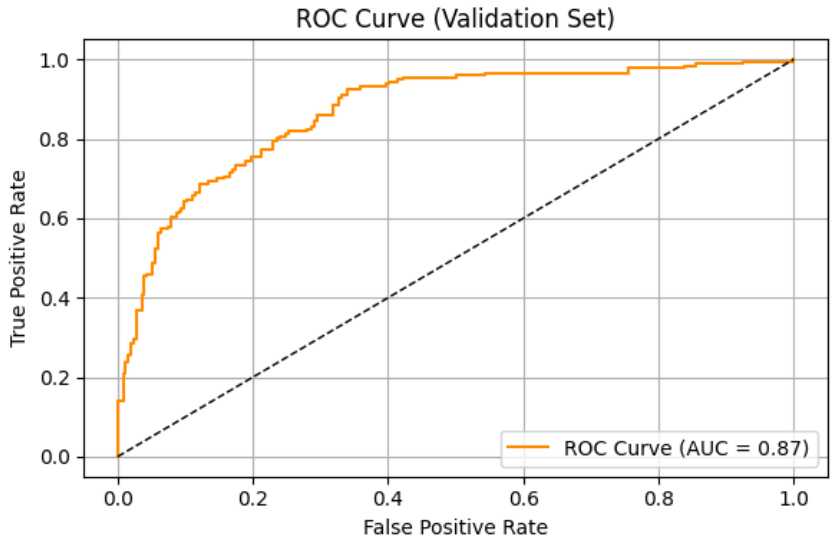


Figure 6: Logistic Regression ROC Curve (Validation Set).

The model's strength lay in identifying No AD cases reliably, but it underperformed in minimizing false negatives—a crucial metric for AD screening.

The confusion matrix (Threshold = 0.3) further highlights this trade-off: while 220 No Alzheimer's cases were correctly predicted, 34 healthy individuals were falsely flagged as AD, and 8 Alzheimer's cases were missed (false negatives). This low number of false negatives is clinically promising but comes at the cost of elevated false positives, which could lead to unnecessary concern or resource usage in real-world deployments.

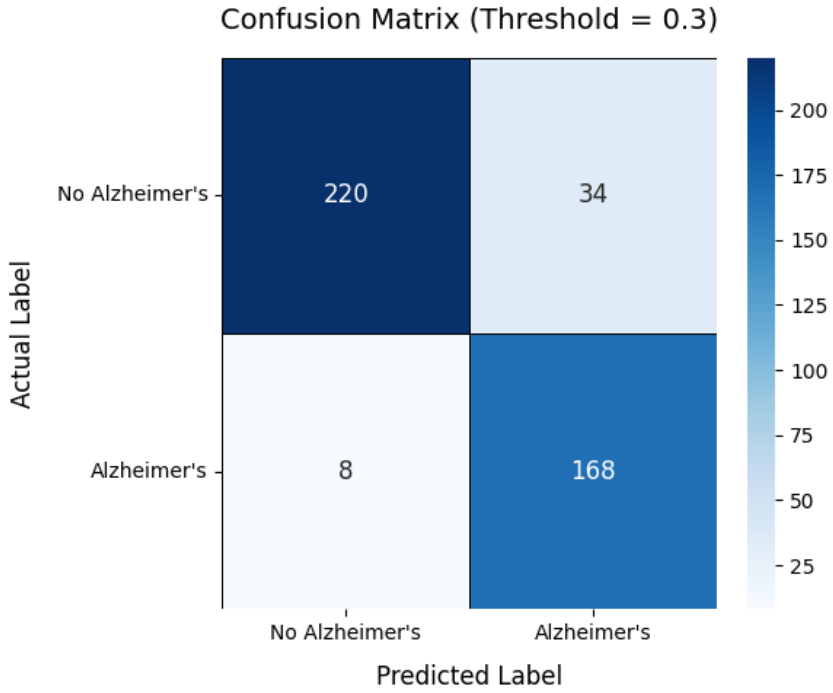


Figure 7: Random Forest Confusion Matrix (Validation Set).

3.4 XGBoost: Final Model

XGBoost emerged as the top-performing model after comprehensive tuning [6]. With `scale_pos_weight = 1.2` and a tuned threshold of 0.45, the model achieved 90.93% accuracy, 0.88 precision, 0.90 recall, and 0.89 F1-score for the AD class on the validation set. This balance demonstrated its superiority over other models in reducing false negatives without excessively increasing false positives.

Confusion Matrix (Validation):

- True Positives (AD): 158
- False Negatives (AD): 18

On the test set, the model maintained 90.77% accuracy and 0.88 F1-score, confirming strong generalization.

3.5 Interpretability and Feature Importance

SHAP analysis revealed that FunctionalAssessment, ADL, MemoryComplaints, and MMSE consistently had the highest predictive influence across both validation and test splits. These features are clinically intuitive, aligning with known Alzheimer's symptoms.

LIME visualizations for individual predictions confirmed the local validity of model decisions, showing that low ADL and high Memory Complaints pushed the model toward AD predictions, while high MMSE and absence of behavioral issues tilted predictions toward No AD.

3.6 Fairness Analysis

We evaluated fairness across gender, age groups, and ethnicities on both the validation and test splits [1].

Gender: On validation data, females had higher recall (94.05%) and F1-score (92.94%) than males (84.78% recall, 87.15% F1), suggesting better sensitivity for female patients. However, on the test set, performance balanced out ($\sim$92% accuracy for both genders), indicating improved generalization.

Age: Validation performance was poor across all age brackets, particularly 70–79 and 90+, with F1-scores below 0.36. Test results showed improved performance, especially for the 80–89 group (F1 = 0.90), although recall remained lower in the 90+ cohort.

Ethnicity: Disparities persisted across splits. While validation results favored African American and Asian patients, test results showed stronger performance for Caucasians and "Other" groups. This inconsistency suggests possible sample imbalance and underscores the need for external validation to ensure equitable deployment.

3.7 Final Model Comparison

Model	Accuracy	Precision (AD)	Recall (AD)	F1 (AD)
Logistic Regression	78.00%	0.71	0.80	0.75
Random Forest	90.23%	0.97	0.78	0.87
XGBoost (Final)	90.93%	0.88	0.90	0.89

Table 1: Final model comparison across evaluation metrics

This comparison clearly shows that XGBoost provides the most balanced and clinically relevant trade-off between sensitivity and overall performance, making it our final choice for deployment.

4 Conclusion and Future Work

In this project, we developed an interpretable and accurate machine learning pipeline to support early prediction of Alzheimer's Disease (AD). Among the models tested, our fine-tuned XGBoost classifier

demonstrated the strongest overall performance, achieving 90.93% accuracy, 0.90 recall, and 0.89 F1-score for the AD class on validation data. Importantly, it retained high generalizability on the unseen test set, reaffirming its potential as a clinically deployable pre-screening tool.

Our results consistently identified Functional Assessment, ADL, Memory Complaints, and MMSE as the most influential features driving Alzheimer's risk—confirming both their predictive power and clinical relevance. SHAP and LIME analyses provided transparency into the model's decisions at both the global and local levels, making the system more trustworthy for real-world use.

Fairness evaluations highlighted promising gender balance on the test set, though ethnic and age-related disparities remained. These insights underscore the need for further fairness-aware tuning and the inclusion of more representative training data.

In future work, we aim to:

- Improve sensitivity further while minimizing false positives
- Evaluate generalization using external clinical datasets
- Implement fairness-aware learning algorithms to reduce subgroup disparities
- Explore ensemble or hybrid interpretability techniques for even more transparent deployment

Overall, our work contributes a robust and interpretable Alzheimer's risk prediction pipeline that bridges machine learning with actionable clinical decision support. If family medicine doctors can make use of this model as an initial-line screening in patients presenting with mild cognitive impairment or complaints of memory problems, it could help flag at-risk individuals even before a formal diagnosis is typically considered. By integrating this risk assessment with AI support into the process for regular check-ups, clinicians are able to diagnose patients with high risks of developing Alzheimer's earlier in the course of the disease. This would allow for early referral of the patients to specialists for comprehensive evaluation, earlier commencement of supportive therapy, access to clinical trials, and practice of lifestyle interventions to slow the disease process. Finally, this model will also have the capability to offer clinicians a data-driven, objective second opinion, leading to active patient care, reduced diagnostic delay, and improved quality of life for patients and families suffering with the ills of Alzheimer's disease.

Our project closely adhered to the original proposal milestones. We successfully implemented and compared three interpretable machine learning models, performed comprehensive SHAP and LIME-based explainability analysis, and validated our final model on both internal validation and external test sets. Fairness evaluations across gender, age, and ethnicity were completed as planned, and we delivered a reproducible pipeline with strong clinical relevance. While we initially planned to test our model on a fully external clinical dataset, our extensive search for suitable open-source datasets revealed that most large-scale Alzheimer's datasets (e.g., ADNI, OASIS-3) were either restricted, required formal data use agreements, or lacked sufficient overlap in features with our working dataset. Despite our best efforts, this made it infeasible to conduct a proper external validation within the project timeline. Nevertheless, all other proposed deliverables were met, and the project outcome remains strongly aligned with our original roadmap [3, 4].

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. 2023. `https://fairmlbook.org/`.

[2] ChatGPT. Prompt 1: Refine introduction section. OpenAI ChatGPT, May 2025. Prompt: "Here is the Introduction section for you to refine to make it sound more professional and comprehensive:...", 2025.

[3] ChatGPT. Prompt 2: Conversion to neurips template. OpenAI ChatGPT, May 2025. Prompt: "Please convert the full final project report draft into NeurIPS template:...", 2025.

[4] ChatGPT. Prompt 3: Remove illogical errors. OpenAI ChatGPT, May 2025. Prompt: "Please go through the entire report draft and remove any illogical redundancies in between sections", 2025.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[6] XGBoost Developers. Xgboost python api documentation. https://xgboost.readthedocs.io/en/stable/python/python_api.html, 2024.

[7] SHAP Documentation. Xgboost with shap — shap docs. https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/XGBoost%20with%20SHAP.html, 2024.

[8] GeeksforGeeks. Managing high-dimensional data in machine learning. https://www.geeksforgeeks.org/managing-high-dimensional-data-in-machine-learning/, 2023.

[9] GeeksforGeeks. Shap: A comprehensive guide to shapley additive explanations. https://www.geeksforgeeks.org/shap-a-comprehensive-guide-to-shapley-additive-explanations/, 2023.

[10] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. Wiley, 2013.

[11] IBM Documentation. Random forest. https://www.ibm.com/think/topics/random-forest, 2023.

[12] Kaggle Contributors. Alzheimer's disease dataset. https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset?resource=download, 2023. Accessed: 2025-05-06.

[13] C. Kavitha et al. Early detection of alzheimer's disease using machine learning. *Frontiers in Public Health*, 2022.

[14] Scikit learn Developers. Scikit-learn: Supervised learning metrics. https://scikit-learn.org/stable/supervised_learning.html, 2024.

[15] Scikit learn Developers. Stratifiedkfold: scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html, 2024.

[16] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[17] Marco Tulio Ribeiro. Lime tabular tutorial. https://marcotcr.github.io/lime/tutorials/Tutorial%20-%20tabular.html, 2024.

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.