Interpretable Alzheimer's Classification Using Gradient Boosting, SHAP and LIME

Group name: ADSA

(Alzheimer's Diagnosis with Sahithi)

Team members: Sahithi C

University of Washington

20250823

I. Introduction

1.1 Research Question

Alzheimer's disease (AD) is the leading cause of dementia , accounting for about 70% of cases worldwide [1]. Despite decades of research, early detection of AD remains a major challenge because symptoms often appear after significant brain damage has already occurred. With advances in imaging and computational methods, researchers now have opportunities to identify hidden disease patterns at earlier stages. This raises the key question: **Can machine learning methods, particularly interpretable models, be used to accurately classify Alzheimer's disease and provide meaningful explanations for clinical use?**

This research question guides our study, as we aim not only to build a predictive model but also to ensure that the results are understandable and trustworthy for clinicians. The goal is to strike a balance between accuracy and interpretability so that the model can support medical decision-making rather than act as a "black box".

1.2 Motivation

Alzheimer's disease has a significant impact on patients, families, and healthcare systems. It damages brain regions responsible for memory, emotions, movement, and daily functioning, gradually leading to complete dependence on caregivers [1]. Traditional diagnostic approaches rely on costly brain imaging procedures, which are not always accessible or feasible in early stages. As a result, many patients are diagnosed late, when treatment options are limited.

Machine learning explores alternate approaches by analyzing patterns in clinical, genetic, or imaging data more efficiently and at lower cost. In particular, interpretable models can highlight which features contribute most to the classification, giving doctors insights into disease progression. By pursuing this approach, we aim to contribute toward cost-effective, early detection tools that could improve patient outcomes and reduce the overall burden of AD on healthcare systems.

1.3 Background (State-of-the-Art)

Alzheimer's was first identified more than a century ago by Dr. Alois Alzheimer in his patient, Auguste Deter, a woman in her 50s who suffered from memory, language, and orientation problems, along with delusional thoughts. After her death, Dr. Alzheimer performed an autopsy and discovered the two hallmark brain features of AD (amyloid plaques and neurofibrillary tangles) [1].

These plaques and tangles lead to neuronal death and the shrinking of brain areas critical for memory, thinking, emotions, movement, and basic skills [1], [2]. Modern imaging technologies now allow

researchers to visualize these pathological changes in living individuals, offering new opportunities to understand the disease and improve diagnosis.

Recent years have seen a surge of interest in applying machine learning to Alzheimer's prediction. Many studies have shown that methods like random forests, support vector machines, and boosting algorithms can achieve high accuracy when applied to clinical, imaging, or biological datasets. However, a major limitation is that many of these models function as "black boxes," making it hard for doctors to interpret how predictions are made.

Our key reference is the work of Kwan et al. (2025) [3], who proposed an explainable gradient boosting approach for Alzheimer's classification. Their study trained three boosting classifiers and used SHAP values to interpret model outputs. The models achieved good performance with an F1 score of ~88%, while remaining transparent, identifying important clinical features that drive predictions. This work demonstrates the feasibility of combining high performance with interpretability and serves as the foundation for our project. Building on this, we aim to reproduce their findings, boosting setup and extend by deepening on the interpretability of Alzheimer's prediction models.

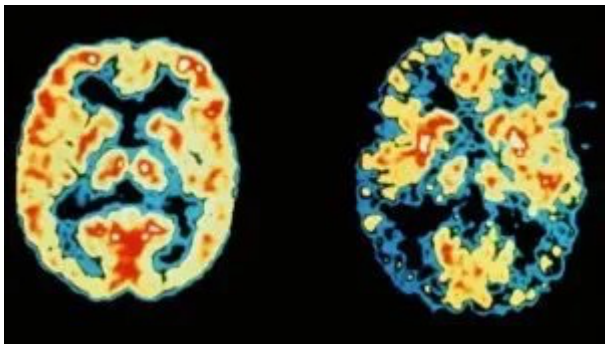


Fig. 1. Brain Imaging Comparison Demonstrating Structural and Volumetric Changes in Neurodegeneration [1]

1.4 Overview of our project

In this project, we investigate **interpretable Alzheimer's classification using gradient boosting**. Using a tabular dataset from Kaggle, we explore demographic, clinical, and imaging-derived features. We designed a sequence of four experiments to carefully analyze the impact of feature selection, class balance, and data leakage on classification outcomes. By structuring the study into multiple stages, we ensure that each step builds on the previous one while addressing potential biases and limitations found in earlier work. We begin by replicating the original study to validate reported findings, then progressively refine the setup. This includes removing the small and noisy converted class, excluding high-leakage features such as MMSE and CDR, and finally improving the recall and reducing false negatives based on the LIME explanations.

II. Related Work

The most related study to our project is by Kwan et al. (2025) [3], which serves as our key reference. They trained three gradient boosting classifiers and applied SHAP values to explain the predictions. Their work identified the most important clinical features for Alzheimer's diagnosis while maintaining strong predictive performance, with F1 score around 88%. A key strength of this paper is its focus on transparency, using interpretable models based on easily available clinical data. This makes their study an ideal foundation for our project, which replicates and extends their approach with deeper interpretability analysis.

Cabanillas-Carbonell and Zapata-Paulini (2025) [4] compared several machine learning models, including Random Forest, AdaBoost, SVM, KNN, and Logistic Regression, on the OASIS dataset that combined both clinical and MRI variables. After pooling two versions of OASIS, they found that Random Forest, SVM, and Logistic Regression achieved very high performance, with accuracy close to 96%. Their study emphasized predictive performance with accessible features but did not place much focus on interpretability, leaving a gap for more explainable models.

Malavika et al. (2020) [5] focused on forecasting Alzheimer's disease using the OASIS longitudinal MRI dataset, which included 150 subjects aged 60–96. They compared several machine learning algorithms such as Logistic Regression, Decision Trees, KNN, SVM, AdaBoost, and Random Forest. Among these, Random Forest performed best with an accuracy of about 86.8%, followed closely by AdaBoost. They also reported sex-stratified performance metrics, showing how male and female patients differed in prediction outcomes. Their main goal was early detection using classic ML techniques, but model transparency was not a priority.

Bari Antor et al. (2021) [6] carried out a comparative study on the OASIS dataset, testing classifiers such as SVM, KNN, Random Forest, and Logistic Regression. Their work showed that Alzheimer's could be detected effectively even on small, publicly available datasets, with competitive accuracy across different models. The study mainly focused on benchmarking the feasibility of these algorithms rather than on interpretability or clinical explanations.

Across these studies, most efforts concentrated on achieving high predictive accuracy, with less emphasis on explainability. Our project builds directly on Kwan et al [3]. by replicating their gradient boosting framework and then extending it with more detailed SHAP and LIME-driven analysis. By doing so, we aim to produce models that are not only accurate but also interpretable, making them more useful for clinical decision-making.

III. Method

3.1 Data

The dataset contains 373 samples. The target variable is Group, indicating the diagnosis category: nondemented, demented, or converted. The dataset includes both demographic and clinical features, such as gender, age, years of education, and socioeconomic status, as well as cognitive and neuroimaging measures like CDR, MMSE, eTIV, nWBV, and ASF.

- **CDR (Clinical Dementia Rating):** Measures the severity of dementia symptoms across memory, orientation, judgment, and daily activities.
- **MMSE (Mini-Mental State Examination):** A cognitive test assessing memory, attention, language, and orientation.
- **eTIV (Estimated Total Intracranial Volume):** Reflects brain size, useful to normalize brain measures across individuals.
- **ASF (Atlas Scaling Factor):** A scaling measure applied to MRI scans, helping adjust for head size in volumetric analysis.
- **nWBV**(**normalized Whole Brain Volume):** Relative to intracranial volume, is the proportion of brain volume relative to the total intracranial volume, used to account for individual head size differences.

These features capture both cognitive performance and brain structure, making them well-suited for machine learning models.

3.2 Exploratory data analysis & feature distribution

Exploratory analysis of the dataset showed that most subjects are nondemented, followed by a smaller number of demented individuals, and an even smaller group of converted cases who may be in transition between nondemented and demented states. The dataset has a fairly balanced gender distribution. The age of participants is mainly between 70 and 85 years, which aligns with the typical age range at higher risk for Alzheimer's disease. Participant's education levels and socioeconomic status vary across the dataset but generally cluster around middle values, indicating a relatively uniform demographic background.

Looking at cognitive and neuroimaging measures, MMSE scores are higher for nondemented subjects, showing better cognitive function, while CDR values are mostly low, reflecting milder dementia symptoms in the population. Structural brain measures such as **eTIV, nWBV, and ASF** follow nearly normal distributions, meaning most values are close to the average with few extreme outliers.

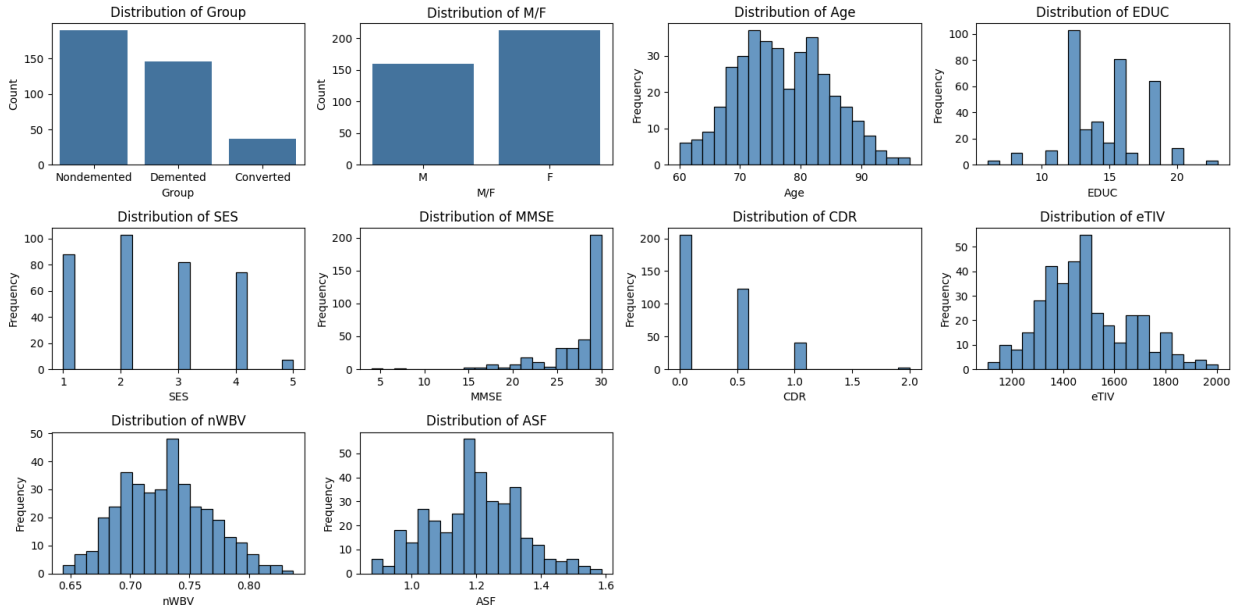


Fig. 2. Exploratory Data Analysis of Key Variables - Demographic, Cognitive, and Brain Imaging Feature Distributions

3.3 Models

This project compares three gradient boosting algorithms for Alzheimer's disease classification: XGBoost, LightGBM, and CatBoost. Gradient boosting methods work by combining multiple weak learners, typically decision trees, into a strong predictive model. Each subsequent tree focuses on correcting the errors of the previous trees, which improves accuracy and reduces bias. These methods are preferred in this project because they are well-suited for tabular datasets and can handle both numerical and categorical features efficiently and finally capture complex relationships between features.

- **XGBoost** is used for its speed and performance, implementing regularization to prevent overfitting.
- **LightGBM** is used to improve training efficiency by using leaf-wise tree growth, to achieve faster computation.
- **CatBoost** is effective with categorical features, automatically handling them without extensive preprocessing and reducing the risk of target leakage.

By comparing these three algorithms, we aim to identify the most efficient model for our dataset while maintaining transparency.

3.4 Evaluation Metrics

To measure how well the models perform, we used four evaluation metrics: Precision, Recall, F1-score, and AUC (Area Under the Curve). These metrics were chosen because the dataset is imbalanced. There are more nondemented cases (190) compared to demented (146) or converted cases (37). Using accuracy alone could be misleading, since a model might predict most subjects as nondemented and still achieve high accuracy without actually identifying dementia cases well.

Precision tells us, out of all the subjects the model predicted as demented, how many were actually correct. Recall measures how many of the actual dementia cases the model was able to detect. Since missing a dementia case is more serious than a false alarm, **Recall is one of our primary focus metrics**.

The F1-score combines precision and recall into a single value, giving a balanced measure when both false positives and false negatives matter. Because of the class imbalance, we also consider the F1-score as a primary focus, as it ensures that neither precision nor recall is ignored.

Finally, the AUC evaluates the model's ability to separate nondemented from demented cases across different thresholds. While it gives an overall sense of performance, our analysis emphasizes **Recall and F1-score** to make sure the models are not just accurate but also reliable for detecting dementia early.

3.5 Interpretable techniques

When using machine learning for healthcare, it is important not only to build accurate models but also to make them **interpretable**. Interpretability means understanding why a model makes certain predictions. In a sensitive domain like dementia detection, doctors and researchers need to trust the model's decisions, so techniques that provide explanations are very valuable.

To achieve this, we used interpretable methods such as **feature importance, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-Agnostic Explanations)**. Feature importance shows which features had the strongest overall effect on predictions. SHAP provides a detailed breakdown of how each feature contributed positively or negatively to a specific prediction. LIME, on the other hand, explains predictions locally by creating a simpler model around a single instance, helping us understand why the model made that decision for a particular patient.

These techniques give both a global and local view of the model. For example, in this project feature importance shows that nWBV (normalized whole brain volume) is the most significant factor overall, while SHAP and LIME explain how a reduced brain volume affects one individual's prediction. This makes the model's decision process more transparent and medically meaningful.

By combining these interpretable techniques, our project ensures the predictions are not just accurate but also **clear, trustworthy, and clinically relevant**, which is essential for real-world healthcare applications.

3.6 Steps of work

Our work started with **data preprocessing**, where we cleaned the dataset by handling missing values, encoding categorical variables, and performing feature selection. For missing values, we used two strategies: **mean imputation for SES** (Skewness = 0.22, close to normal) and **median imputation for MMSE** (Skewness = -2.37, indicating skewness), ensuring that the dataset was consistent and suitable for training.

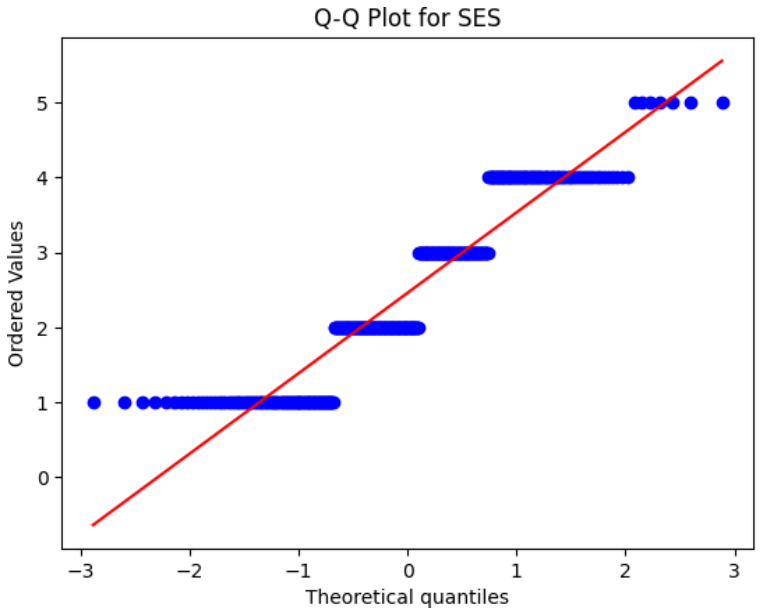


Fig. 3. Statistical Distribution Assessment using Quantile-Quantile (Q-Q) Plots to Determine appropriate Imputation Methods for 'Socio Economic Status (SES)'

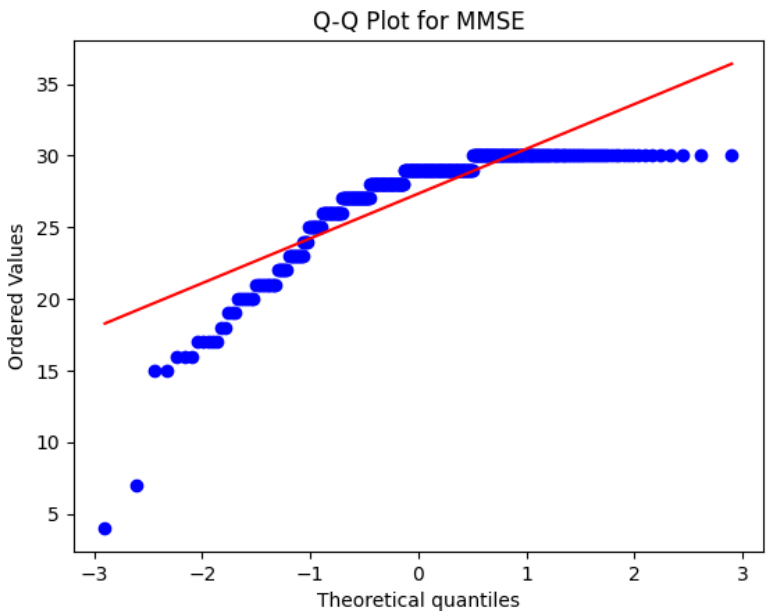


Fig. 4. Statistical Distribution Assessment using Quantile-Quantile (Q-Q) Plots to Determine appropriate Imputation Methods for 'Mini-Mental State Examination (MMSE)'

Next, we performed **model training and evaluation** using three gradient boosting classifiers (XGBoost, CatBoost, LightGBM). Hyperparameters were tuned by following the parameter ranges used in the original paper, though our experiments produced slightly different best parameters. To ensure reliability of results, we used **Repeated Stratified Cross-Validation**, which preserves class distribution across folds and reduces variance from a single random split. Each fold is created with the same ratio of class distributed. This is really useful for small imbalanced datasets because it prevents folds from being skewed toward one class and ensures fairer training and evaluation. Each trained model was then evaluated using Precision, Recall, F1-Score, and AUC, with Recall and F1-Score emphasized to reduce false negatives.

We conducted our work in **four sequential experiments** to systematically improve Alzheimer's classification.

- **Experiment 1:** Focused on establishing baseline results by training the models on all available features, replicating the setup from the original study.
- **Experiment 2**: We removed the "converted" class because it added noise to the dataset. In our data, the converted group contains a mix of CDR values, including 0 (like nondemented cases) and 0.5 – 1 (like demented cases). This overlap makes it difficult for the models to distinguish between nondemented and demented subjects. Removing the converted class created a cleaner binary classification and improved interpretability.
- **Experiment 3**: Refined the previous setup by removing redundant and "cheating" features, such as CDR and MMSE, which directly measure dementia severity and could artificially inflate

performance. This step ensured that the models relied on more generalizable clinical and imaging features.

- **Experiment 4:** Built on Experiment 3 by making small feature adjustments based on insights from interpretability tools like LIME. For example, we removed 'SES' which helped reduce false negatives and improved the model's reliability in identifying demented patients, while keeping the overall performance balanced.

3.7 Process graph

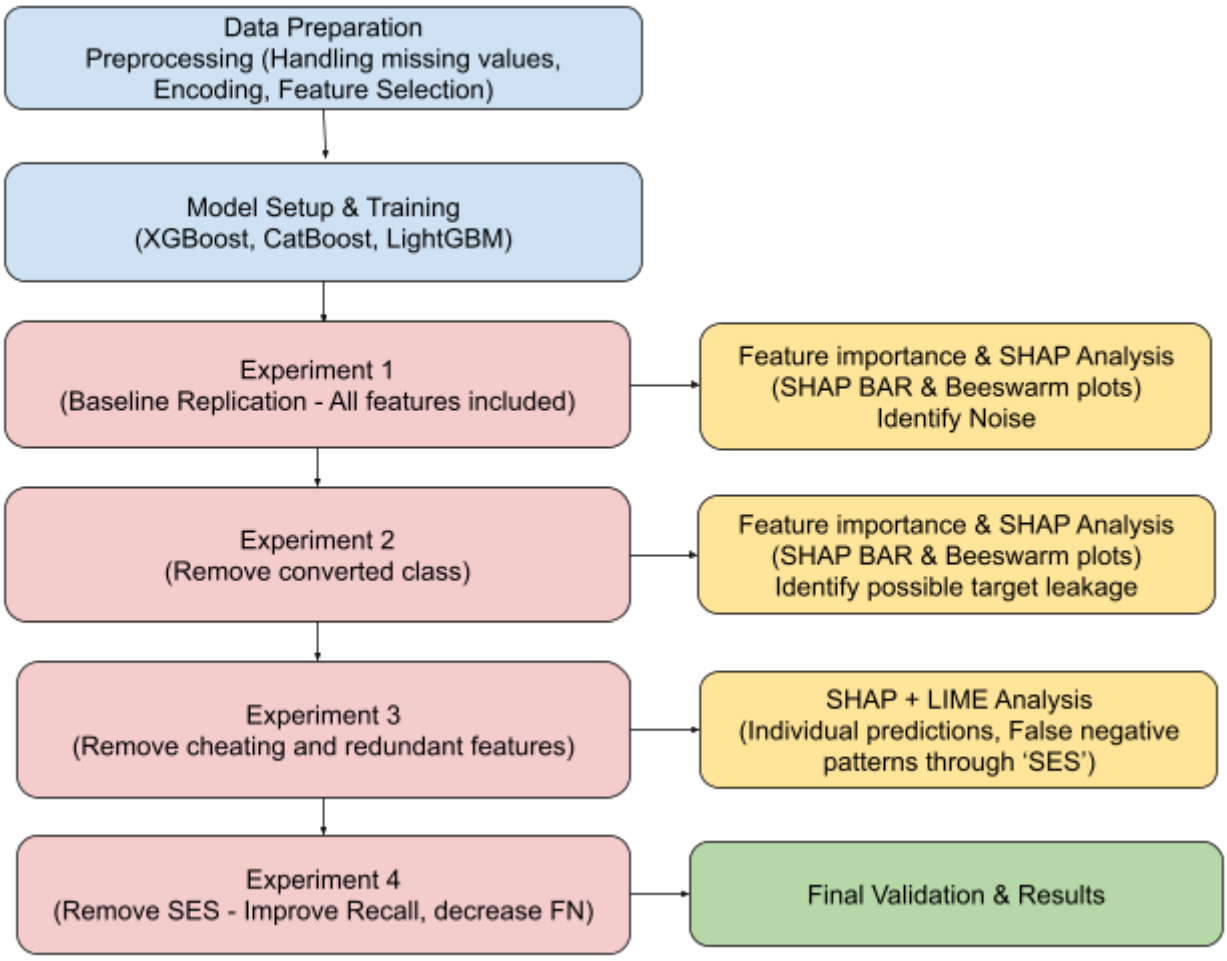


Fig. 5. Iterative Experimental Process Flow Diagram for Interpretable Alzheimer's Classification

IV. Results

4.1 Results for Experiment 1

In Experiment 1, we hypothesized that reproducing the original paper's setup would yield comparable performance to the reported study [3] and that Clinical Dementia Rating (CDR) and Mini-Mental State Examination (MMSE) would emerge as dominant predictors of dementia status. Following the original configuration, the nondemented group was assigned as class 0, while the demented and converted groups were combined as class 1. With this setting, class 0 contained 190 samples and class 1 contained 183 samples.

The hyperparameters for each boosting model in all experiments were fine-tuned using GridSearchCV, with the search space summarized in Table 1. After hyperparameter optimization and repeated stratified cross-validation, model performance was evaluated on a hold-out test set. The results, shown in Table 2, indicate that all three boosting models achieved identical outcomes on the test set, with an accuracy of 0.88, precision of 0.972, recall of 0.814, F1-score of 0.886, and AUC of 0.891. These values are closely aligned with those in the original paper, which reported a precision of 0.8965, recall of 0.88, F1-score of 0.8805, and AUC of 0.9135, though accuracy was not provided.

SHAP analysis was further applied to examine feature contributions. As shown in Fig. 6 (Experiment 1 SHAP bar plots), the Clinical Dementia Rating (CDR) dominated with a disproportionately high importance score relative to other features. CDR is a standardized clinical measure derived from a dementia screening test, commonly used in preliminary assessments of cognitive decline. In both XGBoost and CatBoost models, the Mini-Mental State Examination (MMSE) appeared as the second most important feature. MMSE is a widely used 30-point questionnaire that evaluates key aspects of cognitive function, including orientation, memory, attention, and language, and is frequently used to screen for dementia severity.

Fig. 7 (Experiment 1 SHAP beeswarm plots) further illustrates the influence of these variables, showing that higher CDR values strongly correlated with predictions of Alzheimer's, while patients with a CDR of zero were almost exclusively classified as nondemented. In contrast, other features exhibited more dispersed effects across samples without consistent patterns.

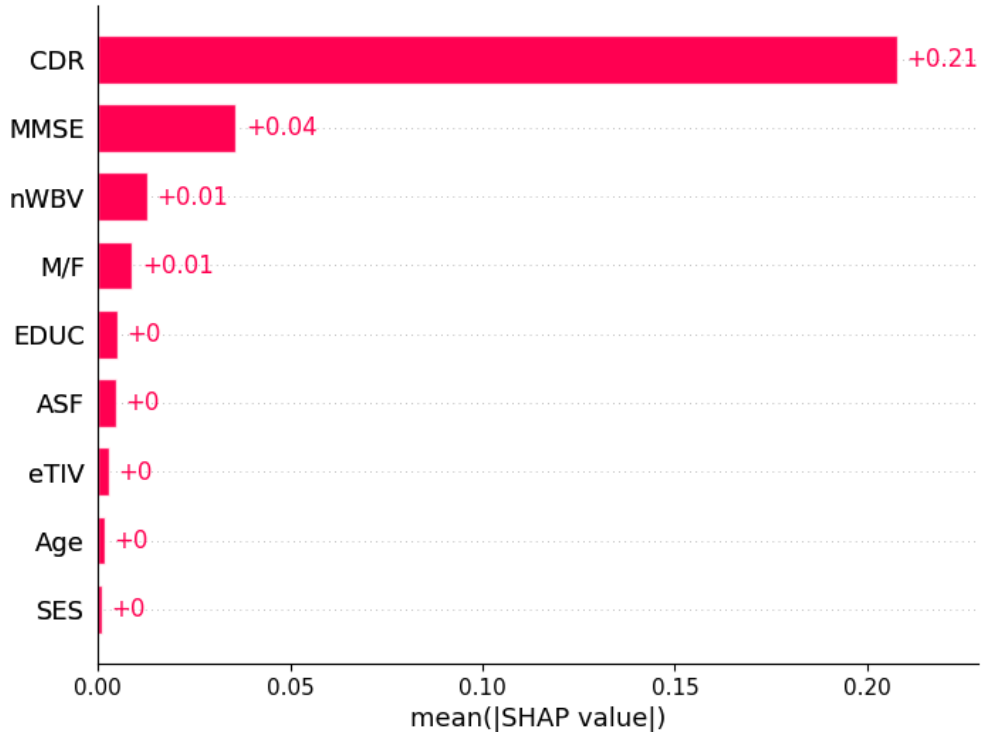
To sum up, Experiment 1 successfully reproduced the original study's results. The findings supported our hypothesis: performance metrics closely matched the original paper, and SHAP confirmed that CDR and MMSE were the strongest predictors of Alzheimer's classification.

Table 1. Hyperparameter Search Space

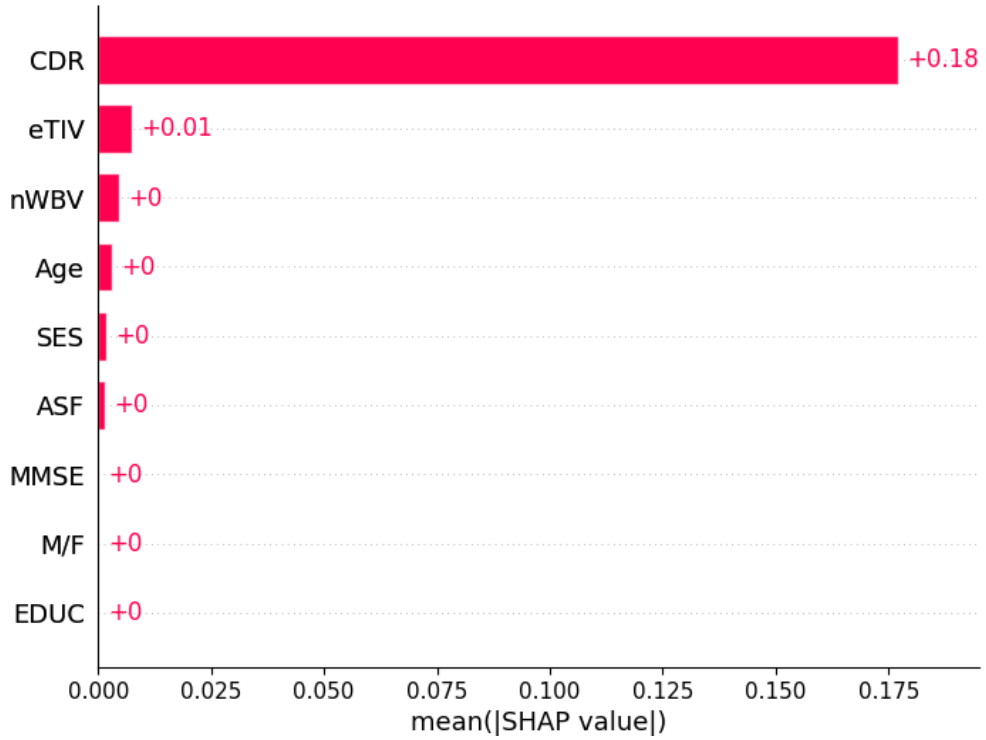
Model	Hyperparameter	Values
XGBoost	n_estimators	100, 200
	max_depth	3, 5, 7
	learning_rate	0.001, 0.01, 0.1
	subsample	0.7, 0.8, 0.9
	colsample_bytree	0.7, 0.8, 1.0
	min_child_weight	1, 3, 5
LightGBM	n_estimators	100, 200
	max_depth	3, 5, 7
	learning_rate	0.001, 0.01, 0.1
	num_leaves	31, 50, 70
CatBoost	min_data_in_leaf	20, 40, 60
	feature_fraction	0.7, 0.8, 1.0
	iterations	100, 200
	depth	3, 5, 7
	learning_rate	0.001, 0.01, 0.1
	l2_leaf_reg	1, 3, 5
	bagging_temperature	0, 0.5, 1

Table 2. Performance result using complete features

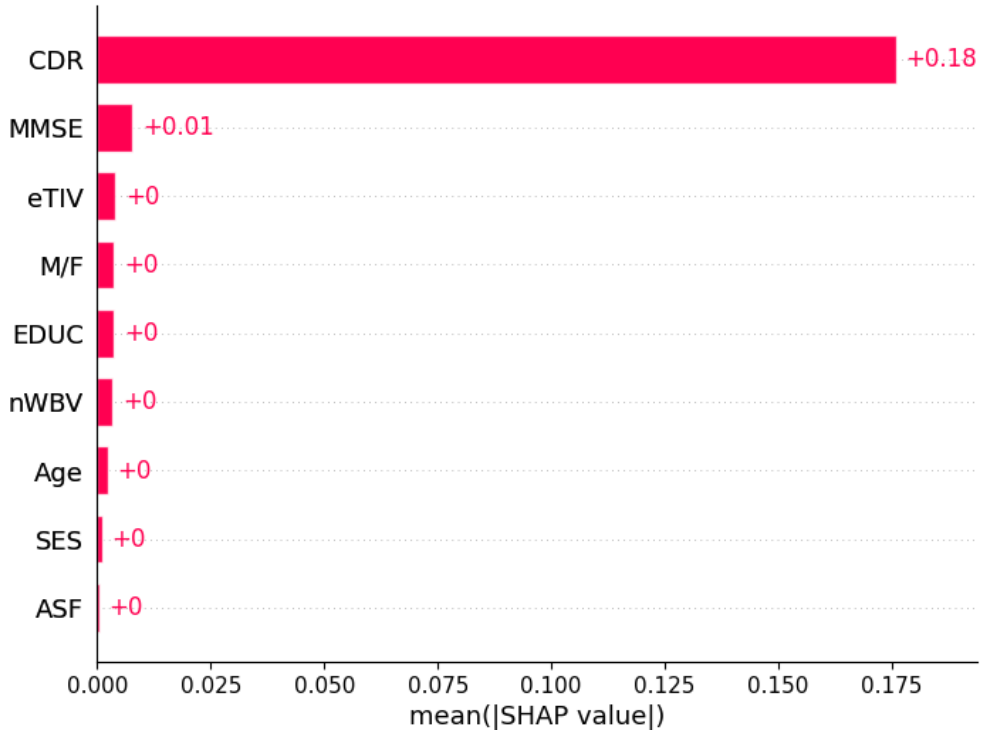
Experiment 1_Repeated Stratified 5×2 CV on TRAIN					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.963± 0.021	0.993 ± 0.014	0.929 ± 0.051	0.959 ± 0.025	0.976 ± 0.018
LightGBM	0.963± 0.020	0.993 ± 0.015	0.929 ± 0.042	0.959 ± 0.023	0.953 ± 0.026
CatBoost	0.963 ± 0.021	0.992 ± 0.015	0.929 ± 0.039	0.959 ± 0.024	0.972 ± 0.021
Experiment 1_Final Performance on HOLD-OUT TEST					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.88	0.972	0.814	0.886	0.891
LightGBM	0.88	0.972	0.814	0.886	0.891
CatBoost	0.88	0.972	0.814	0.886	0.891
Original Paper's Test Performance					
Model		**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost		0.8965	0.88	0.8805	0.9135
LightGBM		0.8965	0.88	0.8805	0.9135
CatBoost		0.8965	0.88	0.8805	0.9135


CDR
MMSE
nWBV
M/F
EDUC
ASF
eTIV
Age
SES
+0.21
+0.04
+0.01
+0.01
+0
+0
+0
+0
+0
0.00
0.05
0.10
0.15
0.20
mean(|SHAP value|)

(a)

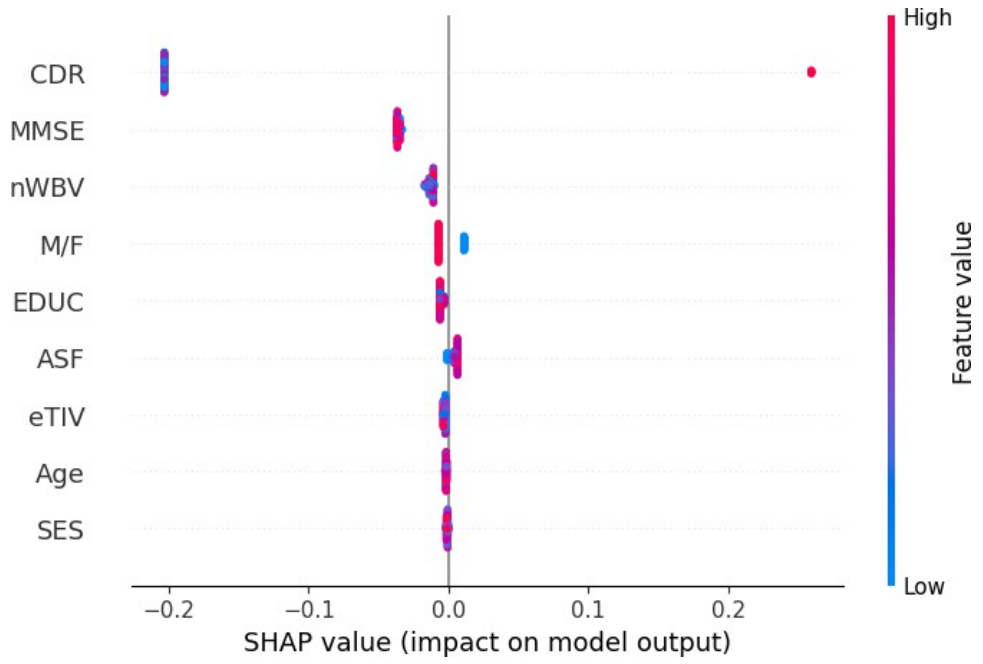

CDR
eTIV
nWBV
Age
SES
ASF
MMSE
M/F
EDUC
+0.18
+0.01
+0
+0
+0
+0
+0
+0
+0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
mean(|SHAP value|)

(b)

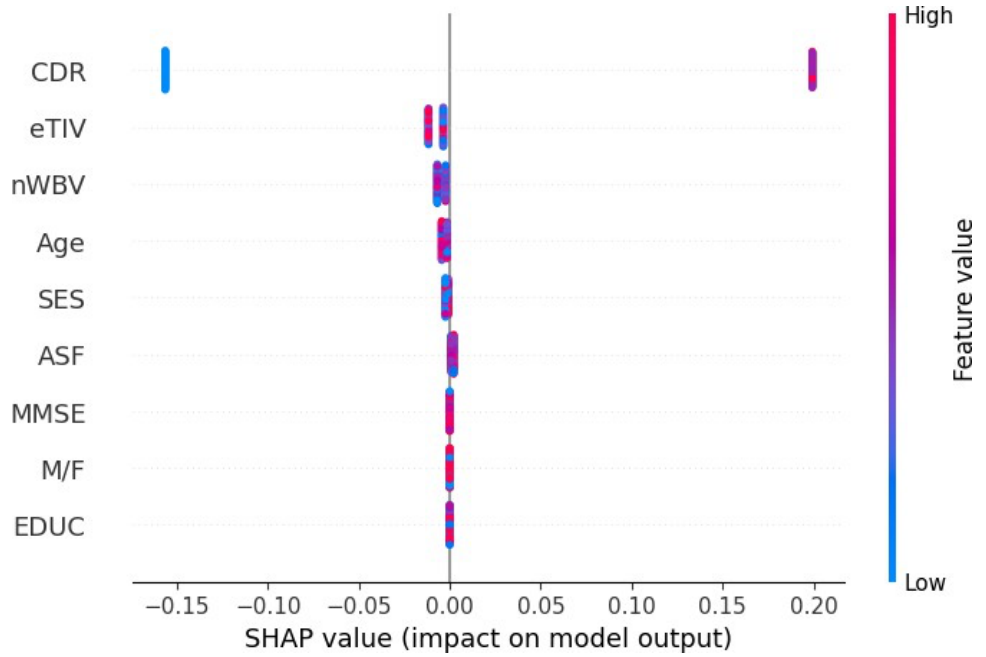


(c)

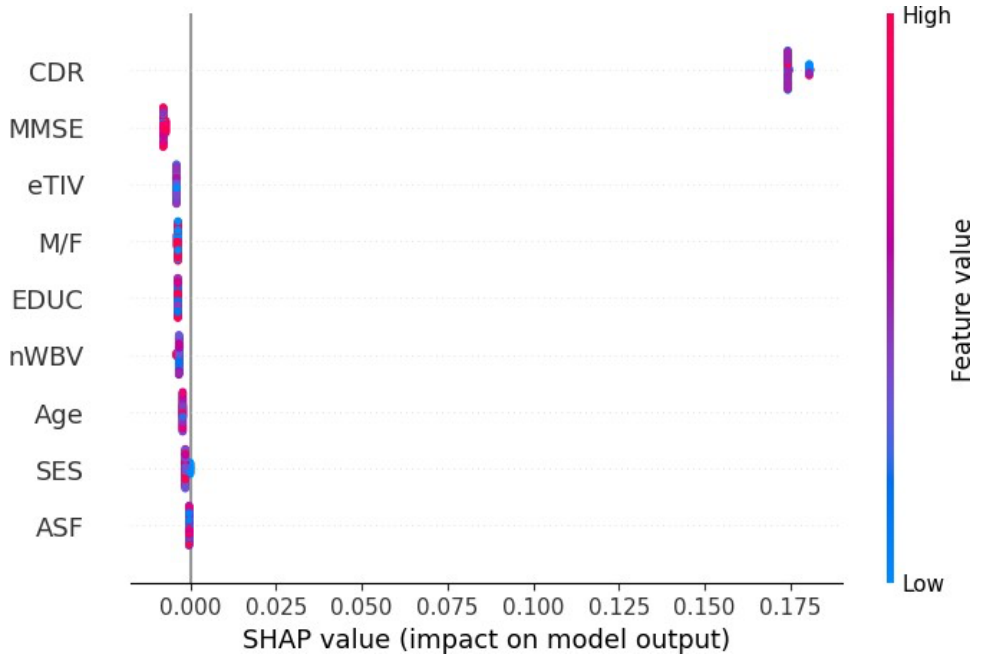
Fig. 6. SHAP Bar Plots for Experiment 1 across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. The plots are generated using the complete feature set and illustrate the feature importance rankings for each model.



(a)



(b)



(c)

Fig. 7. SHAP Beeswarm Plots for Experiment 1 across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. The plots are generated using the complete feature set. Colors represent feature values, ranging from low (blue) to high (red), while the x-axis shows each feature's contribution to the model's predictions. For M/F (gender), red points represent female while blue points represent male.

4.2 Results for Experiment 2

In Experiment 2, we hypothesized that excluding the converted group would yield a cleaner diagnostic boundary between nondemented and demented patients, leading to more reliable classification performance. The converted group refers to patients who were initially nondemented but later diagnosed as demented, making it a mixed and ambiguous category. Prior studies have handled this group inconsistently: some treated it as demented [5], following the same approach as our key reference paper [3] , while others merged it with nondemented [4], and many did not specify their treatment at all [7], [8], [9]. We argued that either inclusion strategy could introduce noise and blur the boundary between nondemented and demented classes.

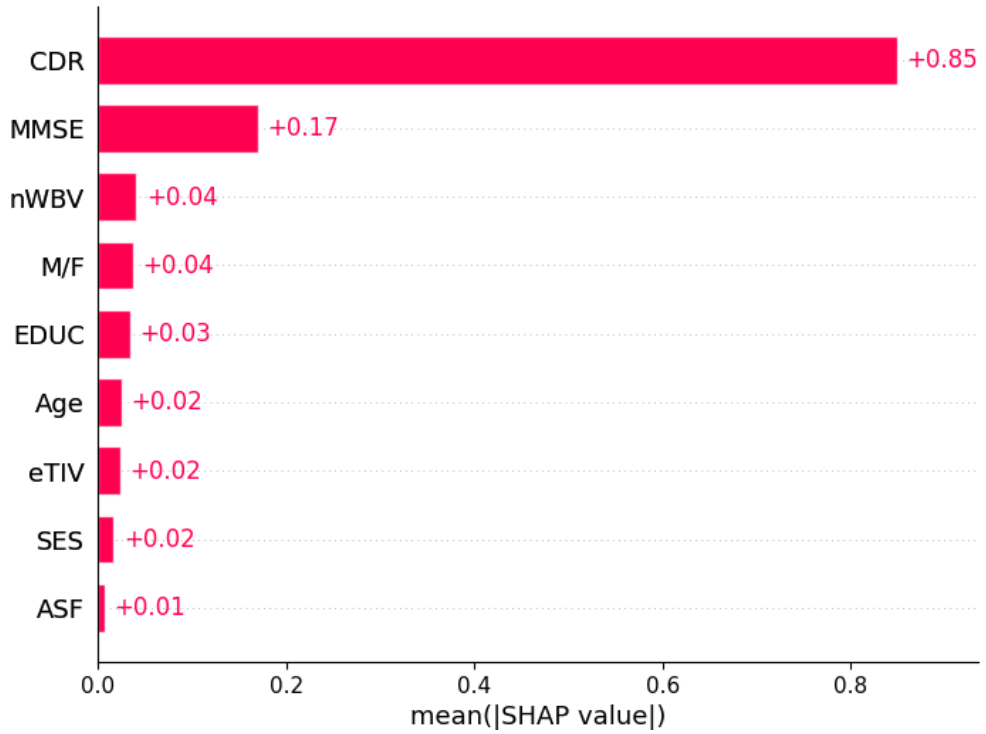
To remove this ambiguity, we excluded the converted group entirely and restricted the task to nondemented (class 0, 190 samples) versus demented (class 1, 146 samples). After hyperparameter optimization and repeated stratified cross-validation, performance was evaluated on a hold-out test set. As shown in Table 3, all three boosting models achieved identical and perfect outcomes, with accuracy, precision, recall, F1-score, and AUC all equal to 1.0.

SHAP analysis, presented in Fig. 8 (bar plots) and Fig. 9 (beeswarm plots), yielded patterns consistent with Experiment 1. CDR remained the most influential predictor, followed by MMSE, while other features showed scattered effects without consistent directionality.

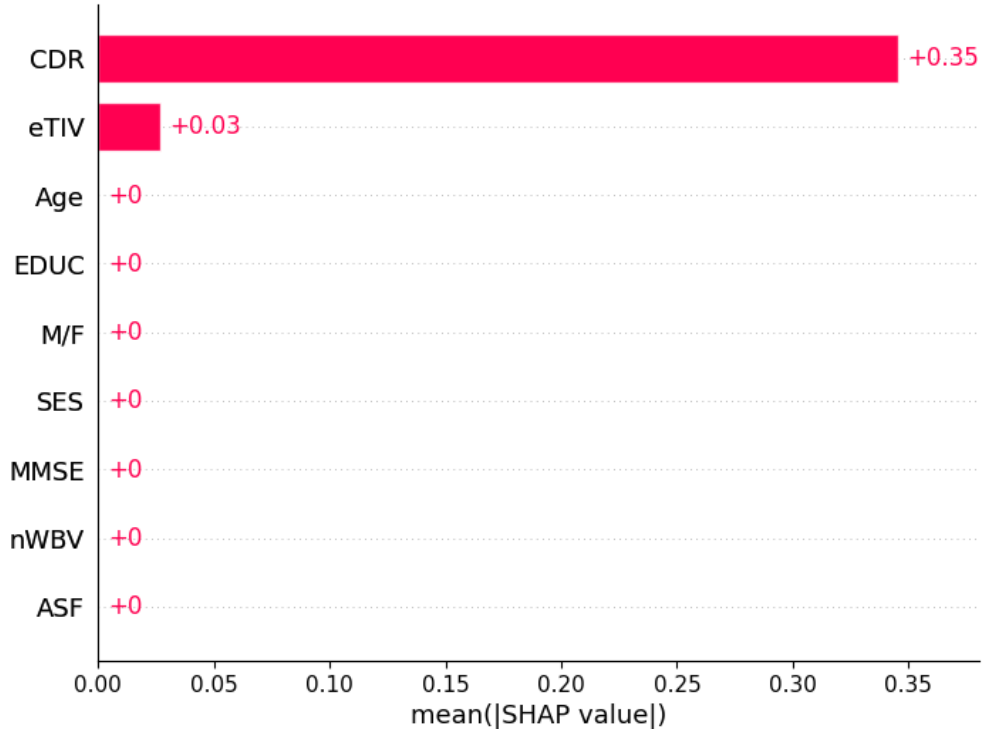
In summary, Experiment 2 supported our hypothesis: excluding the converted group sharpened the diagnostic boundary and dramatically improved performance. However, the perfect results also raised concerns of possible target leakage, suggesting that the observed classification success may not generalize to real-world data.

Table 3. Performance result using complete features without a converted group

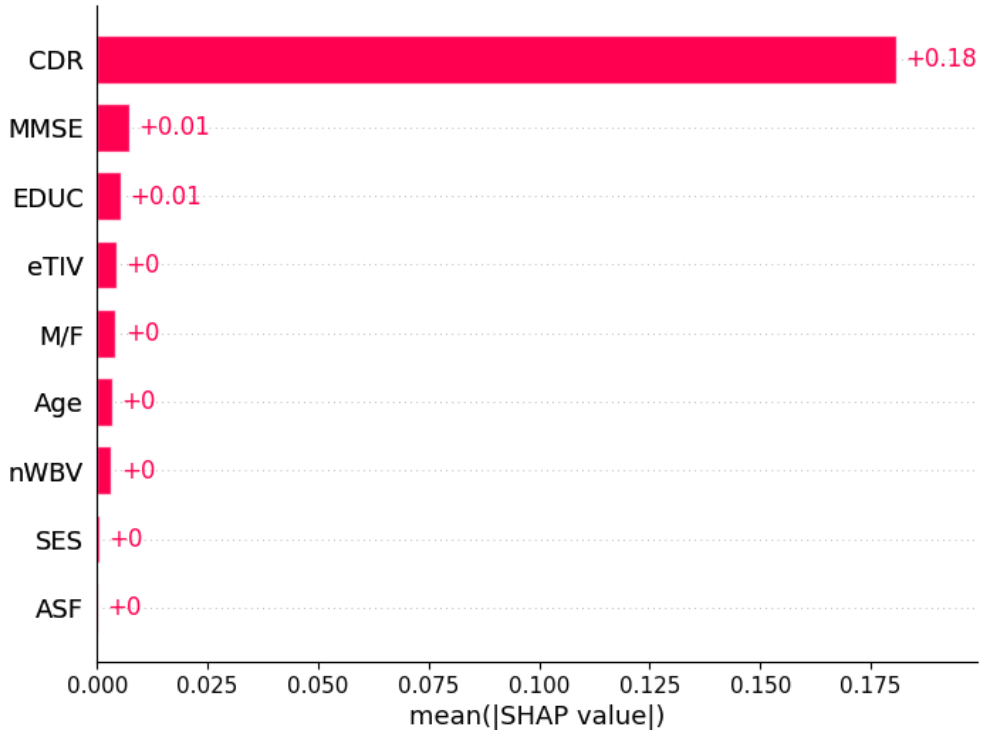
Experiment 2_Repeated Stratified 5×2 CV on TRAIN					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.993± 0.012	0.984 ± 0.027	1.000 ± 0.000	0.992 ± 0.014	0.997 ± 0.004
LightGBM	0.993± 0.009	0.984 ± 0.020	1.000 ± 0.000	0.992 ± 0.010	0.998 ± 0.003
CatBoost	0.993± 0.009	0.984 ± 0.020	1.000 ± 0.000	0.992 ± 0.010	0.995 ± 0.006
Experiment 2_Final Performance on HOLD-OUT TEST					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	1.0	1.0	1.0	1.0	1.0
LightGBM	1.0	1.0	1.0	1.0	1.0
CatBoost	1.0	1.0	1.0	1.0	1.0


CDR
MMSE
nWBV
M/F
EDUC
Age
eTIV
SES
ASF
+0.85
+0.17
+0.04
+0.04
+0.03
+0.02
+0.02
+0.02
+0.01
0.0
0.2
0.4
0.6
0.8
mean(|SHAP value|)

(a)

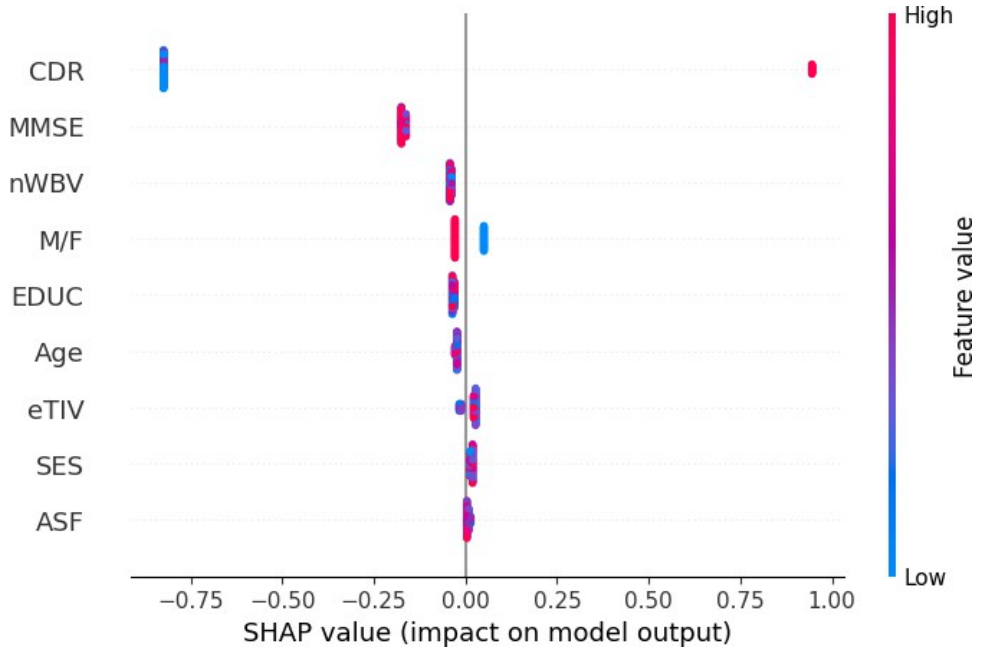

CDR
eTIV
Age
EDUC
M/F
SES
MMSE
nWBV
ASF
+0.35
+0.03
+0
+0
+0
+0
+0
+0
+0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mean(|SHAP value|)

(b)

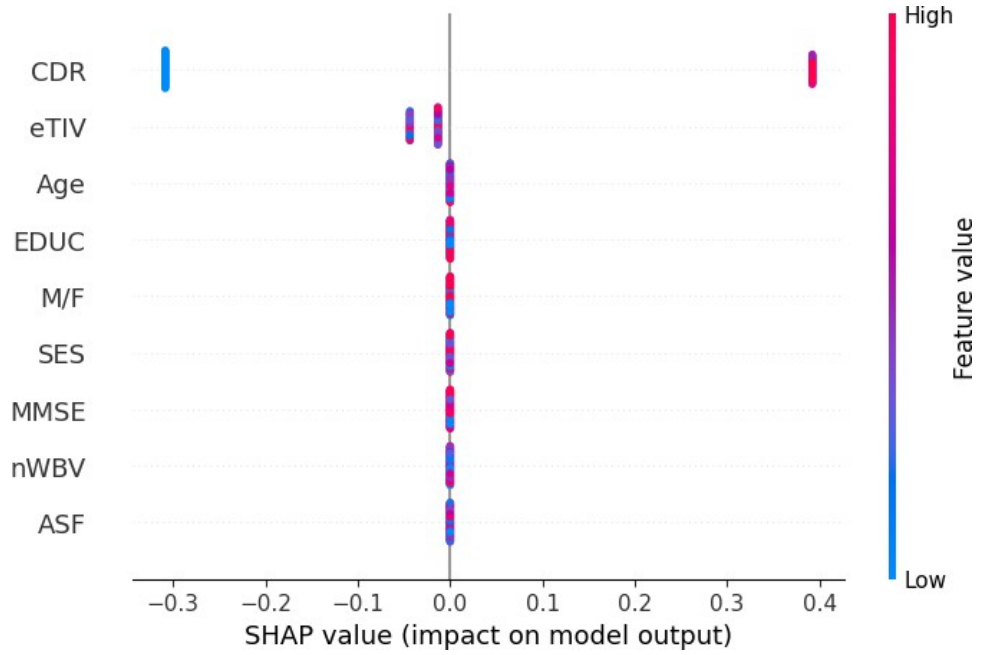


(c)

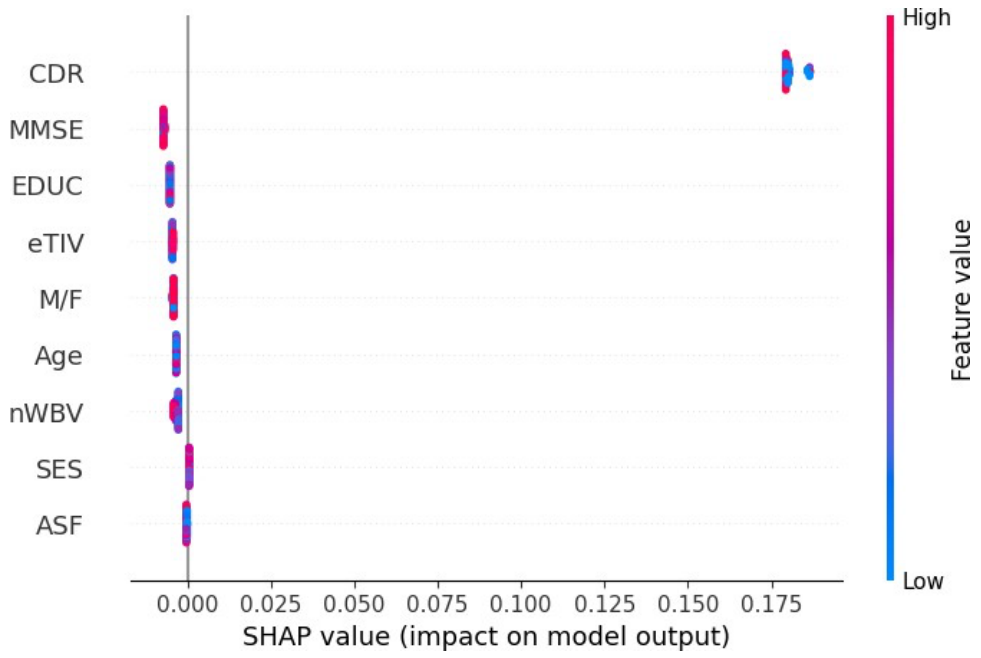
Fig. 8. SHAP Bar Plots for Experiment 2 across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. The plots are generated using the complete feature set without a converted group and illustrate the feature importance rankings for each model.



(a)



(b)



(c)

Fig. 9. SHAP Beeswarm Plots for Experiment 2 across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. The plots are generated using the complete feature set without a converted group. Colors represent feature values, ranging from low (blue) to high (red), while the x-axis shows each feature's contribution to the model's predictions. For M/F (gender), red points represent female while blue points represent male.

4.3 Results for Experiment 3

In Experiment 3, we hypothesized that removing CDR, MMSE, and ASF would eliminate sources of target leakage and redundancy, thereby yielding more modest but realistic performance and more clinically interpretable results.

In the first two experiments, CDR and MMSE were clearly dominating the predictions. The perfect performance in Experiment 2 served as a major red flag, strongly suggesting target leakage and raising concerns about the reliability of the results. To address this issue and obtain a fairer assessment of model performance, in Experiment 3 we first examined the influence of CDR and MMSE individually to evaluate whether they were indeed the primary sources of target leakage.

We began with CDR. As shown in Fig. 10, the distribution of CDR by group reveals that all samples with CDR = 0 belong to the nondemented class, while nearly all samples with non-zero values belong to the demented class. When CDR values are binarized into 'zero' versus 'non-zero' and compared against the ground-truth labels, the confusion matrix in Fig. 11 shows only 0.6% false positives and overall accuracy close to 100%. Similarly, a single-feature model trained solely on CDR achieved nearly perfect scores, as reported in Table 4. This demonstrates that CDR alone can almost fully determine the class label, confirming it as a classic case of target leakage. Including this feature in classification is essentially equivalent to cheating, and therefore it must be removed from further experiments.

Next, we examined MMSE. As shown in Fig. 12, the distribution of MMSE scores by group reveals a clear boundary: all samples with an MMSE score below 25 fall into the demented class. When used as a single predictor, MMSE alone achieved approximately 91% accuracy, as reported in Table 4. This strong predictive power indicates that MMSE, like CDR, acts as a source of target leakage. Consequently, MMSE must also be removed in order to ensure a fair and unbiased evaluation of model performance.

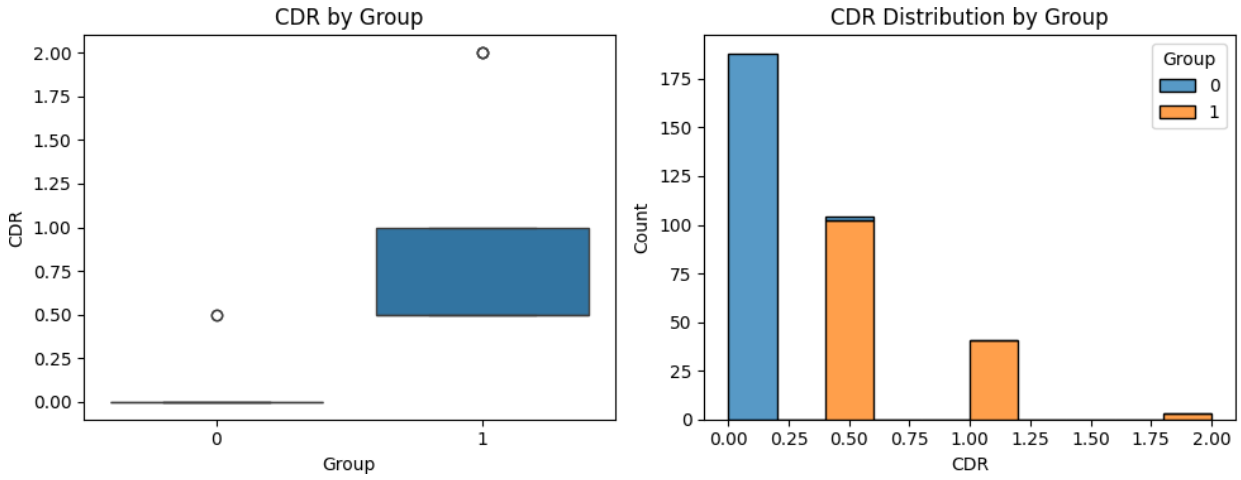


Fig. 10. CDR distribution by group (box plot and hist plot), group 0 (blue) represents nondemented and group 1 (orange) represents demented.

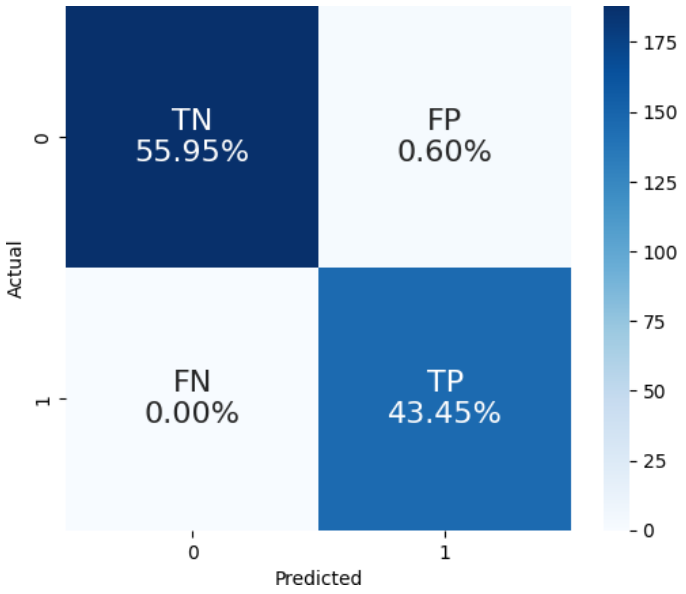


Fig. 11. Confusion matrix of binarized CDR values (zero vs. non-zero) compared with ground-truth group labels.

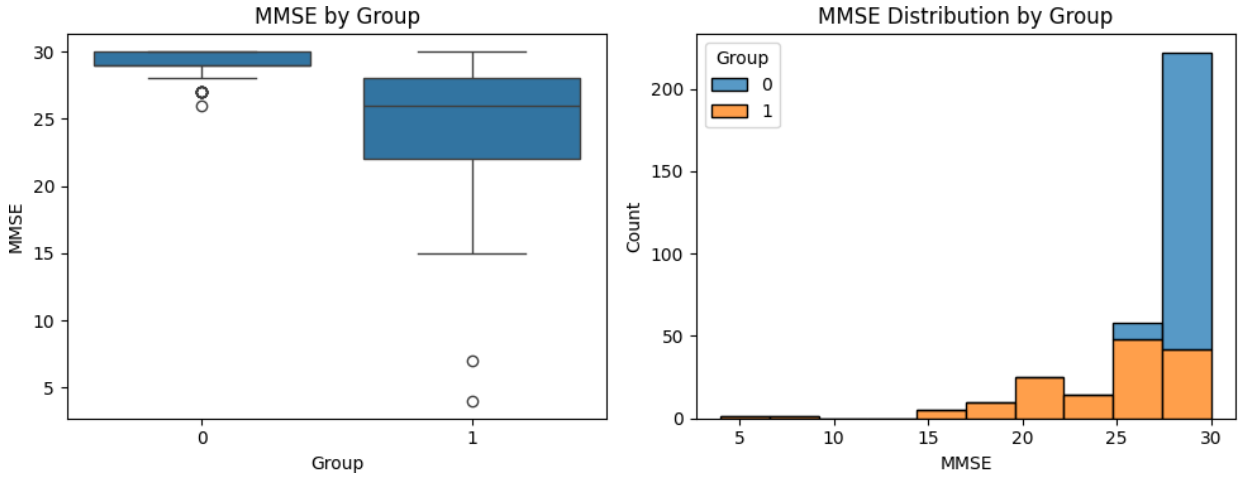


Fig. 12. MMSE distribution by group (box plot and hist plot), group 0 (blue) represents nondemented and group 1 (orange) represents demented.

Table 4. Single feature Modeling Performance

Single feature Random Forest Modeling Performance (weighted average)				
Feature	**Accuracy**	**Precision**	**Recall**	**F1**
CDR only	1.00	1.00	1.00	1.00
MMSE only	0.91	0.92	0.91	0.91

In addition to leakage, we identified redundancy issues. As shown in Fig. 13, the correlation heatmap reveals that eTIV (Estimated Total Intracranial Volume) and ASF (Atlas Scaling Factor) are almost perfectly correlated, with a coefficient of 0.99. This relationship is biologically reasonable, since eTIV is approximately equal to the Template Intracranial Volume multiplied by ASF [10]. Including both features in the model therefore introduces redundancy without adding meaningful information. To reduce multicollinearity and simplify the feature set, we removed ASF from subsequent experiments.

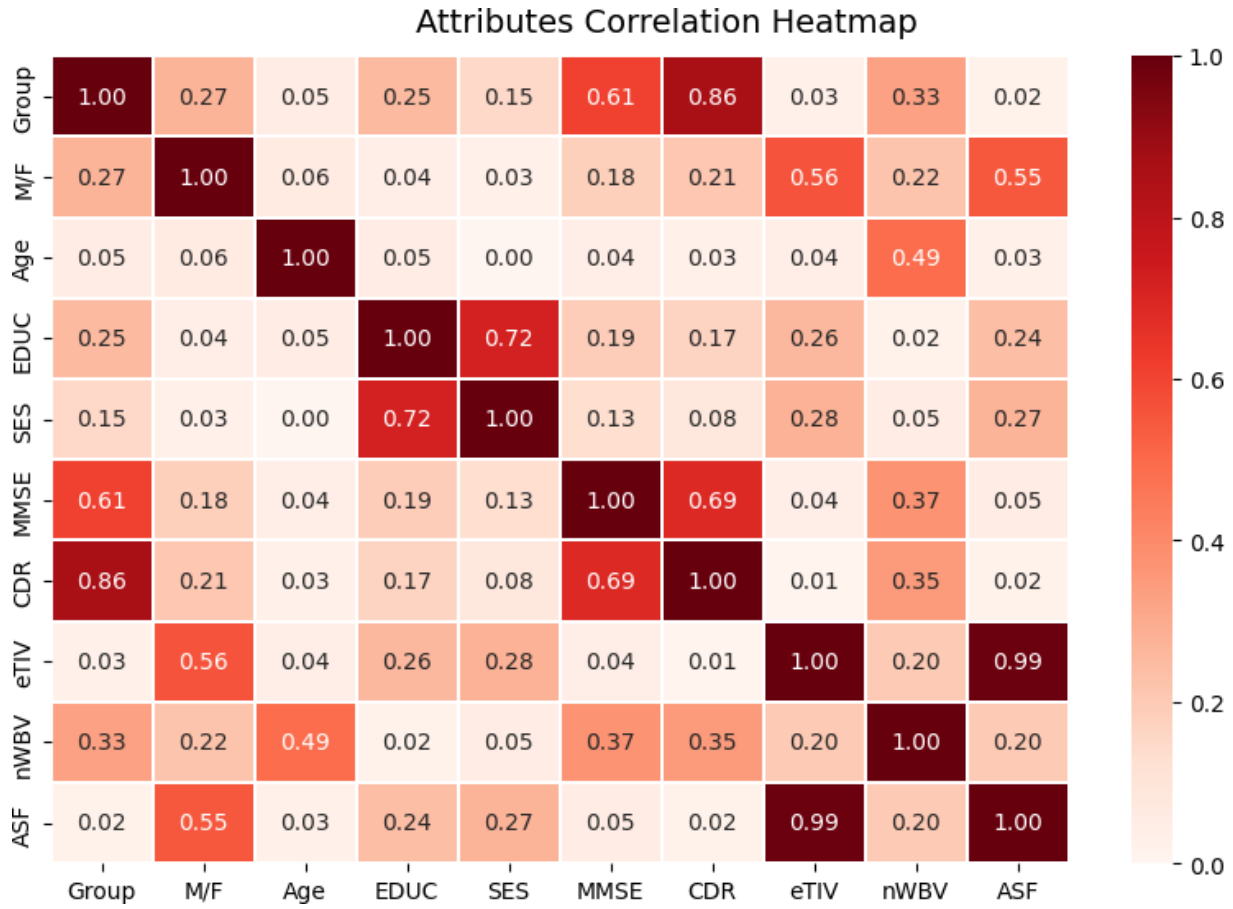


Fig. 13. Correlation heatmap of all features, illustrating pairwise relationships and potential redundancies in the dataset.

With leakage and redundancy addressed, we proceeded to Experiment 3 using a cleaned feature set of six variables (excluding CDR, MMSE, and ASF). The objective of Experiment 3 was to prevent models from relying on diagnostic variables directly tied to the ground-truth labels and instead evaluate how well the remaining features could discriminate between nondemented (class 0) and demented (class 1) patients. After hyperparameter tuning and repeated stratified cross-validation, model performance was evaluated on a hold-out test set. As shown in Table 5 and Fig.14, performance dropped compared to Experiments 1 and 2, confirming the strong influence of the removed features. However, the results remained meaningful, as they reflect a more realistic evaluation without target leakage. Among the

three boosting models, CatBoost consistently achieved the best balance, with an accuracy of 0.8676, Precision of 0.8387, Recall of 0.8667, F1-score of 0.8525, and ROC_AUC of 0.9289 on the hold-out test. XGBoost followed closely as a stable middle ground, while LightGBM showed greater variability, with a larger gap between training and test performance. Importantly, even without the diagnostic "cheating" features, all models maintained strong predictive ability, with recall above 80%.

Table 5. Performance result with feature selection (removing CDR/MMSE/ASF)

Experiment 3_Repeated Stratified 5×2 CV on TRAIN					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.8005 ± 0.0613	0.7818 ± 0.0930	0.7638 ± 0.0690	0.7696 ± 0.0647	0.8551 ± 0.0533
LightGBM	0.7783 ± 0.0724	0.7604 ± 0.1032	0.7335 ± 0.0954	0.7416 ± 0.0783	0.8376 ± 0.0706
CatBoost	0.8267 ± 0.0677	0.8023 ± 0.0864	0.8025 ± 0.0837	0.8007 ± 0.0756	0.8841 ± 0.0492
Experiment 3_Final Performance on HOLD-OUT TEST					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.8382	0.7879	0.8667	0.8254	0.8860
LightGBM	0.7647	0.7188	0.7667	0.7419	0.8658
CatBoost	0.8676	0.8387	0.8667	0.8525	0.9289

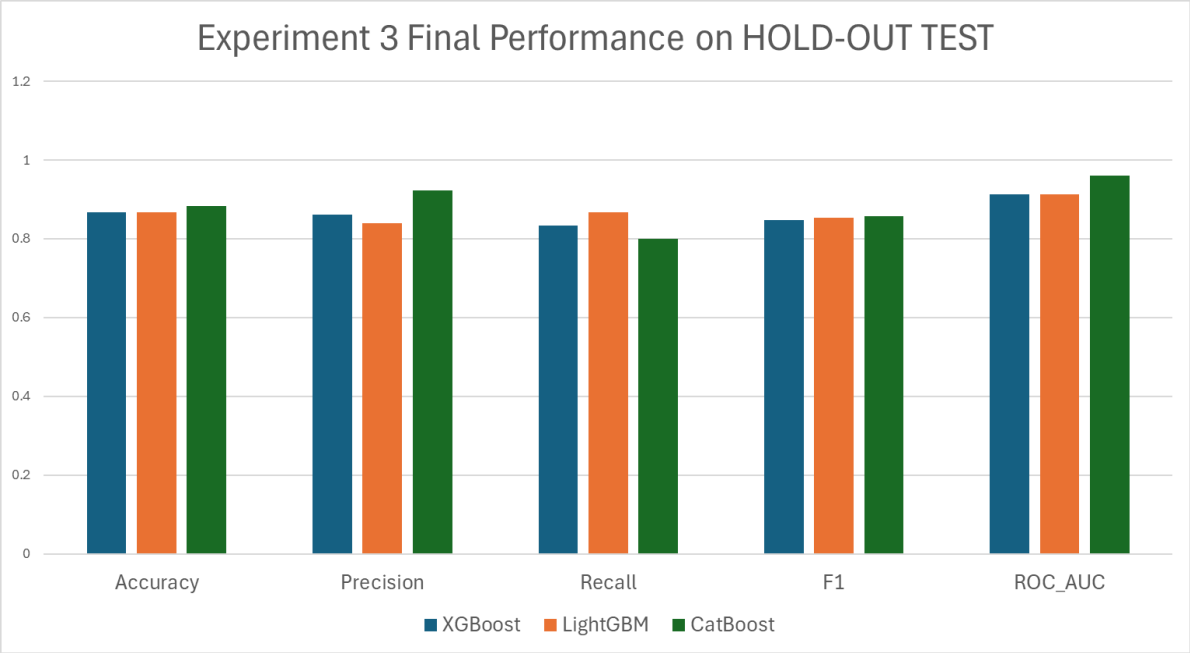
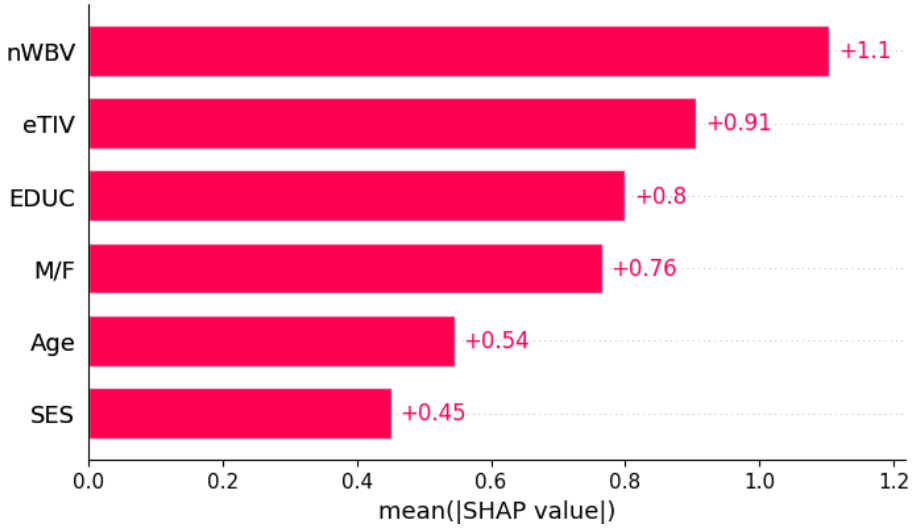


Fig. 14. Experiment 3_Performance result on Hold-out Test with feature selection (removing CDR/MMSE/ASF)

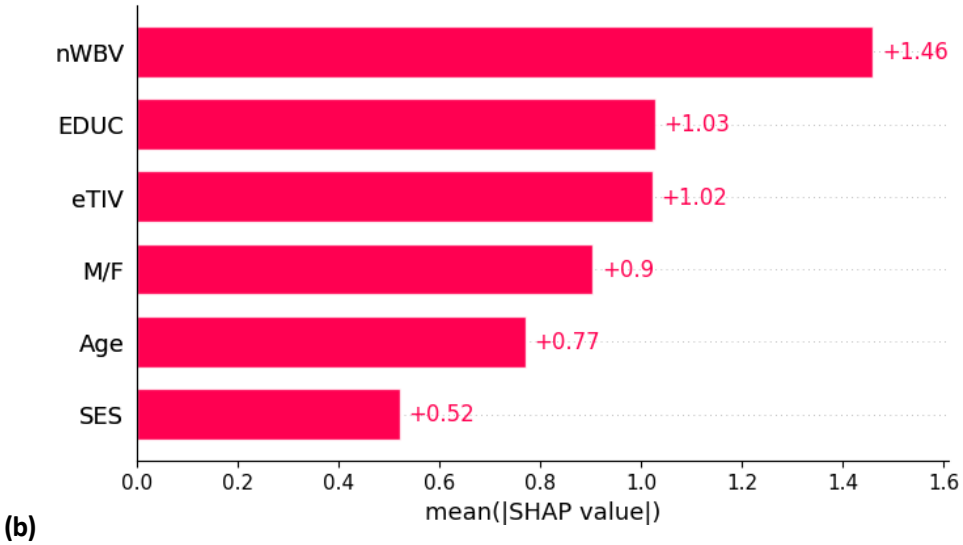
With the cleaned feature set, the SHAP results became more clinically meaningful. As shown in Fig. 15 (Experiment 3 SHAP bar plots), all three models exhibited similar patterns: normalized whole brain volume (nWBV) emerged as the most important feature, while socioeconomic status (SES) contributed the least. Fig. 16 (Experiment 3 SHAP beeswarm plots) provides further insight, showing that higher values of nWBV, eTIV (estimated total intracranial volume), and education consistently pushed predictions toward the nondemented class.

These findings align well with established domain knowledge and prior research. Lower nWBV and eTIV values reflect greater brain atrophy, which has been strongly associated with cognitive decline and dementia progression [3]. Similarly, higher education levels are linked to a reduced risk of Alzheimer's disease, as they are associated with greater cognitive reserve, a protective factor that helps maintain cognitive function despite underlying neurodegeneration [11].

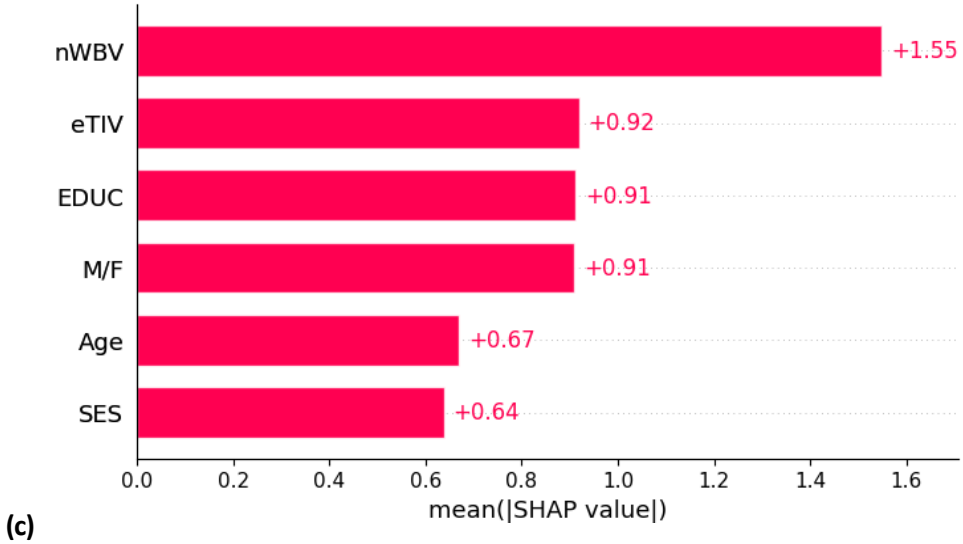
By contrast, SES showed no clear or consistent pattern, with mixed contributions across samples. For age and M/F (gender), SHAP indicated that older age and female sex pushed predictions toward the nondemented class. This result is counterintuitive, as clinically age is a strong risk factor for Alzheimer's disease [12], and women are often overrepresented among AD patients due to their longer life expectancy [13]. However, some studies have reported no significant gender difference once age is controlled [14]. Because SHAP revealed this unexpected pattern, we conducted further analyses focusing specifically on age and gender.



(a)

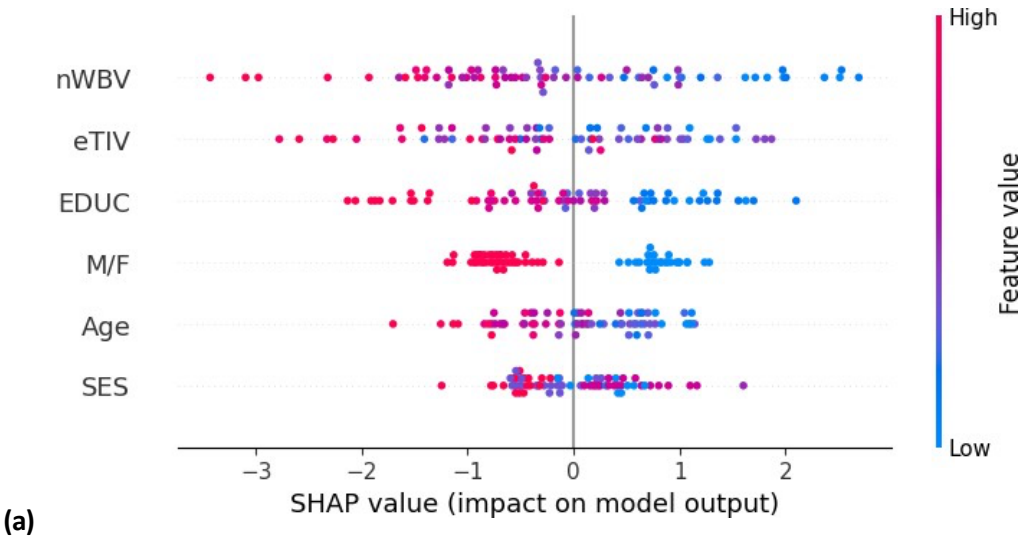


(b)

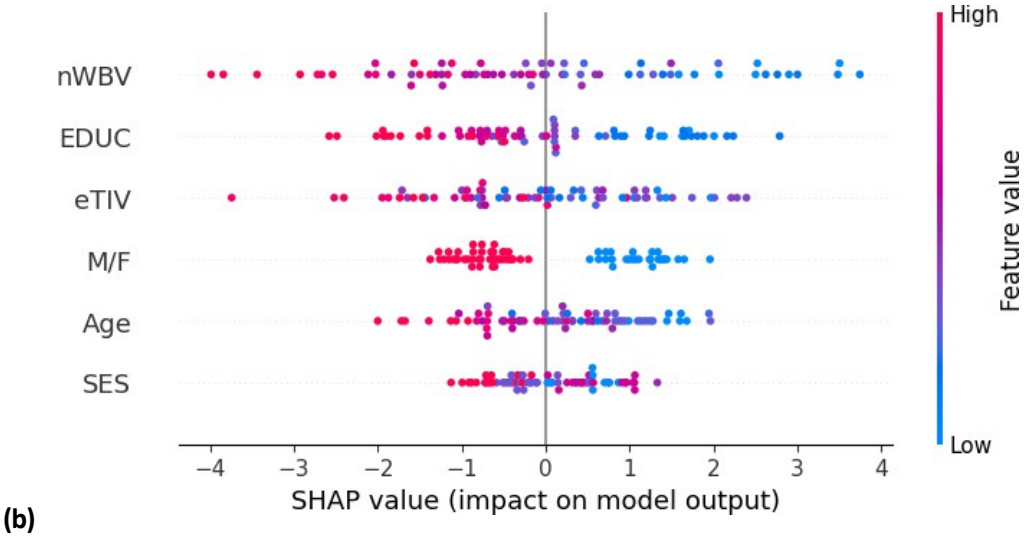


(c)

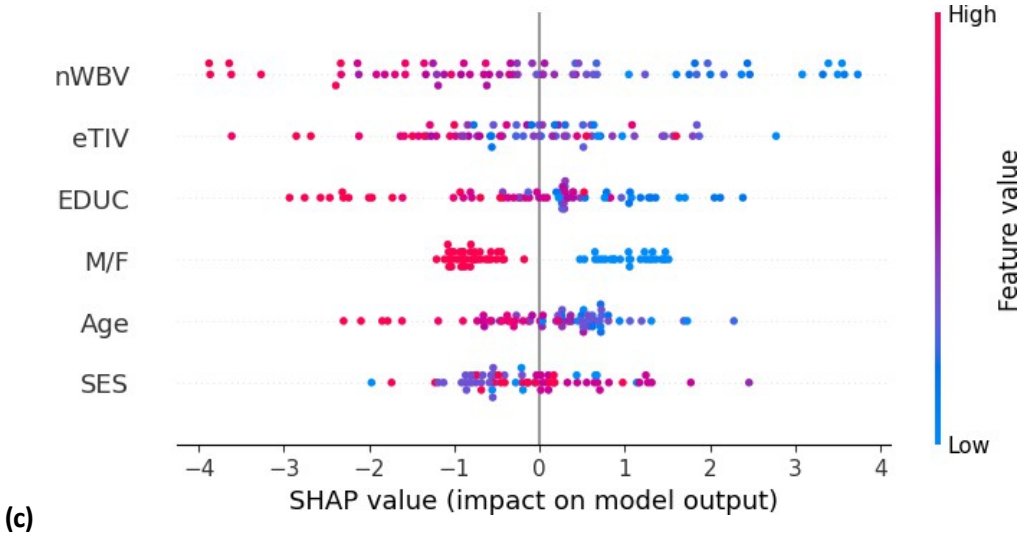
Fig. 15. SHAP Bar Plots for Experiment 3 across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. The plots are generated with feature selection (removing CDR/MMSE/ASF) and illustrate the feature importance rankings for each model.


nWBV
eTIV
EDUC
M/F
Age
SES
−3
−2
−1
0
1
2
SHAP value (impact on model output)
High
Feature value
Low

(a)


nWBV
EDUC
eTIV
M/F
Age
SES
−4
−3
−2
−1
0
1
2
3
4
SHAP value (impact on model output)
High
Feature value
Low

(b)


nWBV
eTIV
EDUC
M/F
Age
SES
−4
−3
−2
−1
0
1
2
3
4
SHAP value (impact on model output)
High
Feature value
Low

(c)

Fig. 16. SHAP Beeswarm Plots for Experiment 3 across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. The plots are generated with feature selection (removing CDR/MMSE/ASF). Colors represent feature values, ranging from low (blue) to high (red), while the x-axis shows each feature's contribution to the model's predictions. For M/F (gender), red points represent female while blue points represent male.

In addition to SHAP values, we also examined SHAP interaction values for age and gender. As shown in Fig. 17, the patterns differ by sex: for males (blue points), older age increases the likelihood of being predicted as demented, whereas for females (red points), older age unexpectedly pushes predictions toward the nondemented class. This discrepancy may reflect biological factors—for example, men experiencing stronger brain atrophy at older ages[15]—or it may simply result from data imbalance. The next step is to investigate whether such imbalance exists in our dataset.

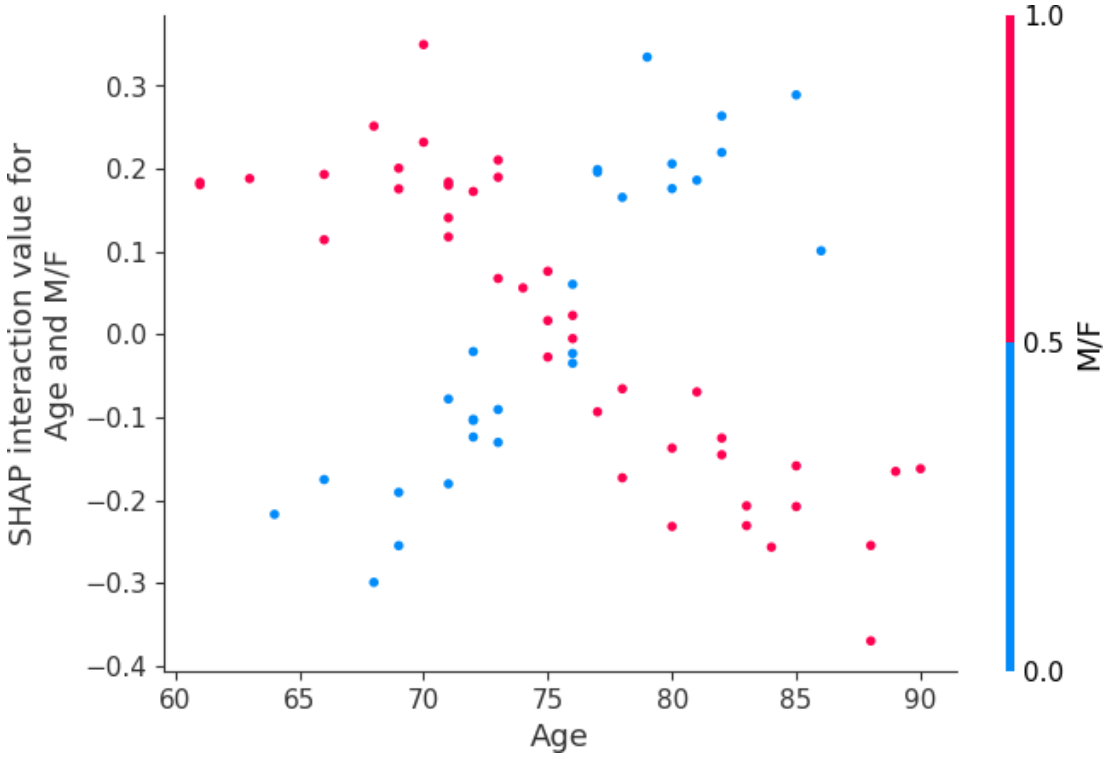


Fig. 17. SHAP interaction value for Age and Gender, blue points represent males while red points represent females.

As shown in Fig. 18 the gender distribution reveals an imbalance: most individuals in the nondemented group are female, while the majority in the demented group are male. This confirms the presence of gender imbalance in the dataset. Age, by contrast, is approximately normally distributed across both groups. However, when age is divided into bands (<70, 70–80, >80), as shown in Fig. 19, an important pattern emerges. In the 80+ band, nearly 60% of participants are female, and many of them belong to the nondemented group. At the same time, the proportion of demented patients within this age band

decreases. Taken together, these demographic imbalances help explain why the SHAP analysis revealed the counterintuitive pattern of older females being pushed toward the nondemented class.

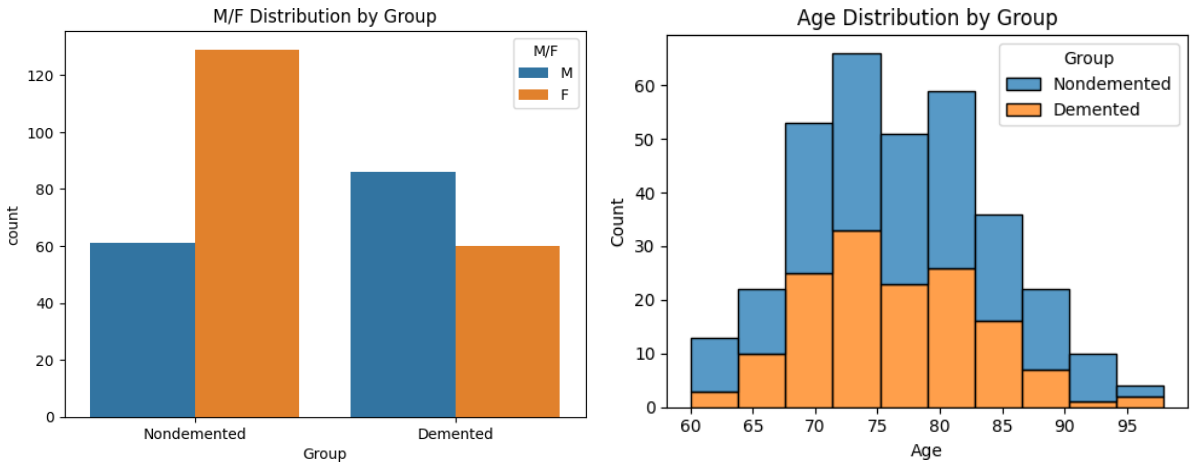
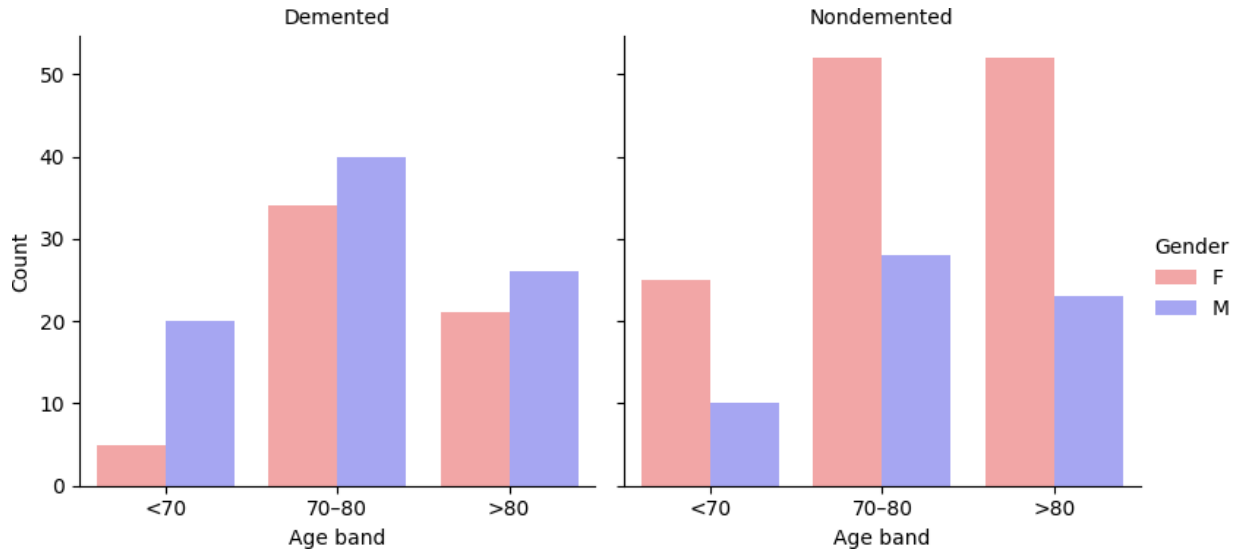


Fig. 18. (left) Gender distribution by group, (right) Age distribution by group.



AgeBand	n_total	n_demented	n_female	pct_demented	pct_female
<70	60	25	30	41.7	50.0
70–80	154	74	86	48.1	55.8
>80	122	47	73	38.5	59.8

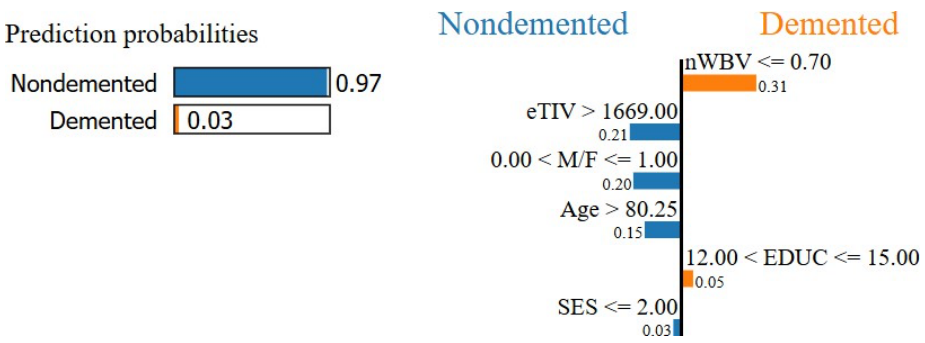
Fig. 19. Age-band and gender distribution of demented and nondemented groups.

To further investigate how such patterns affect model reliability, we turned our attention to individual misclassifications. In Alzheimer's classification, false negatives are particularly concerning, since missing a true patient is far more harmful than mistakenly flagging a healthy individual. Using LIME, we analyzed

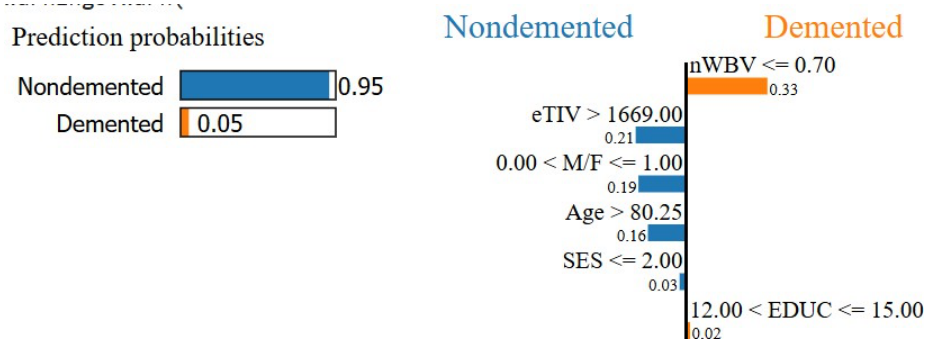
individual prediction failures and identified a common false negative case, index 31, that all three models misclassified. This patient was demented but consistently predicted as nondemented. As shown in Fig. 20, the LIME explanations illustrate why: the patient exhibited features typically associated with nondemented status, being an older female with high eTIV and low SES, which misled all models.

We then expanded the LIME analysis across all false negatives to see which features most often pushed these patients toward the non-demented class. As shown in Fig. 21, a recurring pattern emerged: SES frequently contributed to false negatives among three models. Since SHAP global analysis already ranked SES as the least important feature, we expected that removing SES might slightly reduce overall performance, but could also help lower false negatives and improve recall. This led directly to Experiment 4, where we evaluated the impact of excluding SES on model performance and false negatives reduction.

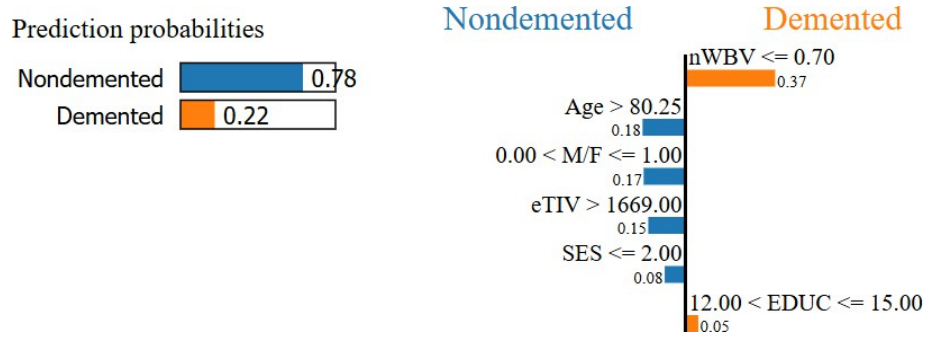
To sum up, Experiment 3 confirmed that removing CDR, MMSE, and ASF eliminated sources of target leakage and redundancy, yielding more modest but realistic performance. The cleaned feature set produced results that were not only stable but also clinically interpretable, with nWBV, eTIV, and education emerging as meaningful predictors. At the same time, Experiment 3 highlighted the challenges posed by demographic imbalances, such as gender and age, which led to counterintuitive patterns. The results from interpretability analysis also motivated the design of Experiment 4, where we examined whether removing SES could further improve model recall. Overall, the findings supported our hypothesis: the models became less inflated by leakage, more realistic, and clinically meaningful, though also revealed dataset imbalance challenges.



(a)

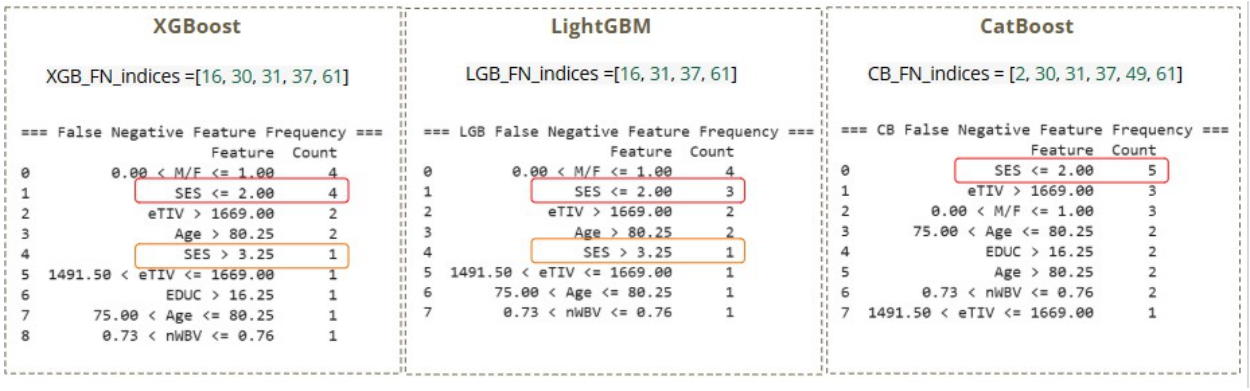


(b)



(c)

Fig. 20. Lime Explanation for Experiment 3's common False Negative case (X_test.iloc[31]) across the three boosting algorithms: (a) XGBoost, (b) LightGBM, and (c) CatBoost. For M/F (gender), 1 represents female while 0 represents male.



XGBoost

XGB_FN_indices =[16, 30, 31, 37, 61]

=== False Negative Feature Frequency ===

	Feature	Count
0	0.00 < M/F <= 1.00	4
1	SES <= 2.00	4
2	eTIV > 1669.00	2
3	Age > 80.25	2
4	SES > 3.25	1
5	1491.50 < eTIV <= 1669.00	1
6	EDUC > 16.25	1
7	75.00 < Age <= 80.25	1
8	0.73 < nWBV <= 0.76	1

LightGBM

LGB_FN_indices =[16, 31, 37, 61]

=== LGB False Negative Feature Frequency ===

	Feature	Count
0	0.00 < M/F <= 1.00	4
1	SES <= 2.00	3
2	eTIV > 1669.00	2
3	Age > 80.25	2
4	SES > 3.25	1
5	1491.50 < eTIV <= 1669.00	1
6	75.00 < Age <= 80.25	1
7	0.73 < nWBV <= 0.76	1

CatBoost

CB_FN_indices = [2, 30, 31, 37, 49, 61]

=== CB False Negative Feature Frequency ===

	Feature	Count
0	SES <= 2.00	5
1	eTIV > 1669.00	3
2	0.00 < M/F <= 1.00	3
3	75.00 < Age <= 80.25	2
4	EDUC > 16.25	2
5	Age > 80.25	2
6	0.73 < nWBV <= 0.76	2
7	1491.50 < eTIV <= 1669.00	1

Fig. 21. False Negative feature frequency across the three boosting algorithms

4.4 Results for Experiment 4

In Experiment 4, we hypothesized that removing SES would improve model recall and reduce false negatives, as suggested by the interpretability analysis in Experiment 3. To test this, we removed SES and re-evaluated performance.

As shown in Fig.22 and Table 6, recall improved noticeably for XGBoost (from 83% to nearly 87%) and CatBoost (from 80% to nearly 87%), while other metrics remained at acceptable levels with only minor drops. LightGBM, however, performed worse after SES was removed. In Experiment 3, we had already observed LightGBM's instability. Although SES did not appear more important in LightGBM than in the other models according to SHAP values and feature rankings, its removal had a disproportionate effect. This may relate to LightGBM's tree-splitting strategy, which can overfit to weaker or noisy features, making the model more sensitive to their absence.

We also compared confusion matrices between Experiments 3 and 4. As shown in Fig. 23, XGBoost reduced false negatives from 5 to 4, while CatBoost reduced false negatives from 6 to 4. LightGBM, however, showed no improvement.

To sum up, Experiment 4 demonstrated that removing SES helped improve recall and reduce false negatives for XGBoost and CatBoost, confirming the value of interpretability-guided feature selection. LightGBM, by contrast, remained unstable and did not benefit from this adjustment. Overall, the findings supported our hypothesis for XGBoost and CatBoost, but not for LightGBM.

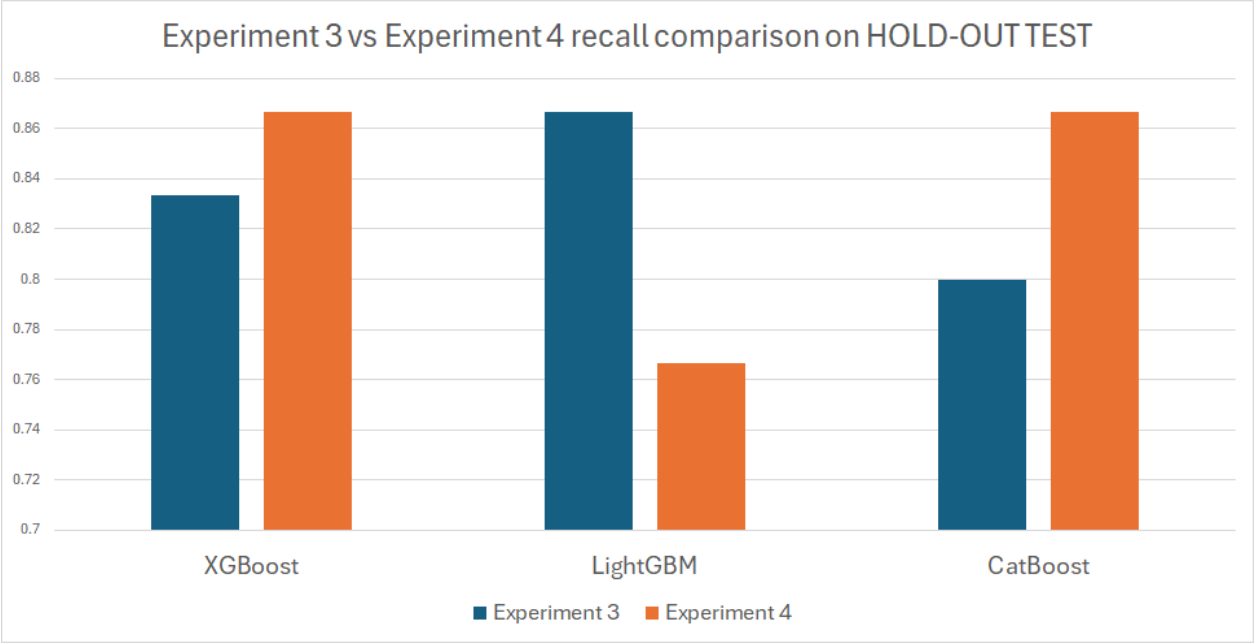


Fig. 22. Experiment 3 vs Experiment 4 (remove SES based on Experiment 3) recall comparison

Table 6 Performance result with feature selection (removing CDR/MMSE/ASF/SES)

Experiment 4_Repeated Stratified 5×2 CV on TRAIN					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.8005 ± 0.0613	0.7818 ± 0.0930	0.7638 ± 0.0690	0.7696 ± 0.0647	0.8551 ± 0.0533
LightGBM	0.7783 ± 0.0724	0.7604 ± 0.1032	0.7335 ± 0.0954	0.7416 ± 0.0783	0.8376 ± 0.0706
CatBoost	0.8267 ± 0.0677	0.8023 ± 0.0864	0.8025 ± 0.0837	0.8007 ± 0.0756	0.8841 ± 0.0492
Experiment 4_Final Performance on HOLD-OUT TEST					
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC_AUC**
XGBoost	0.8382	0.7879	0.8667	0.8254	0.8860
LightGBM	0.7647	0.7188	0.7667	0.7419	0.8658
CatBoost	0.8676	0.8387	0.8667	0.8525	0.9289

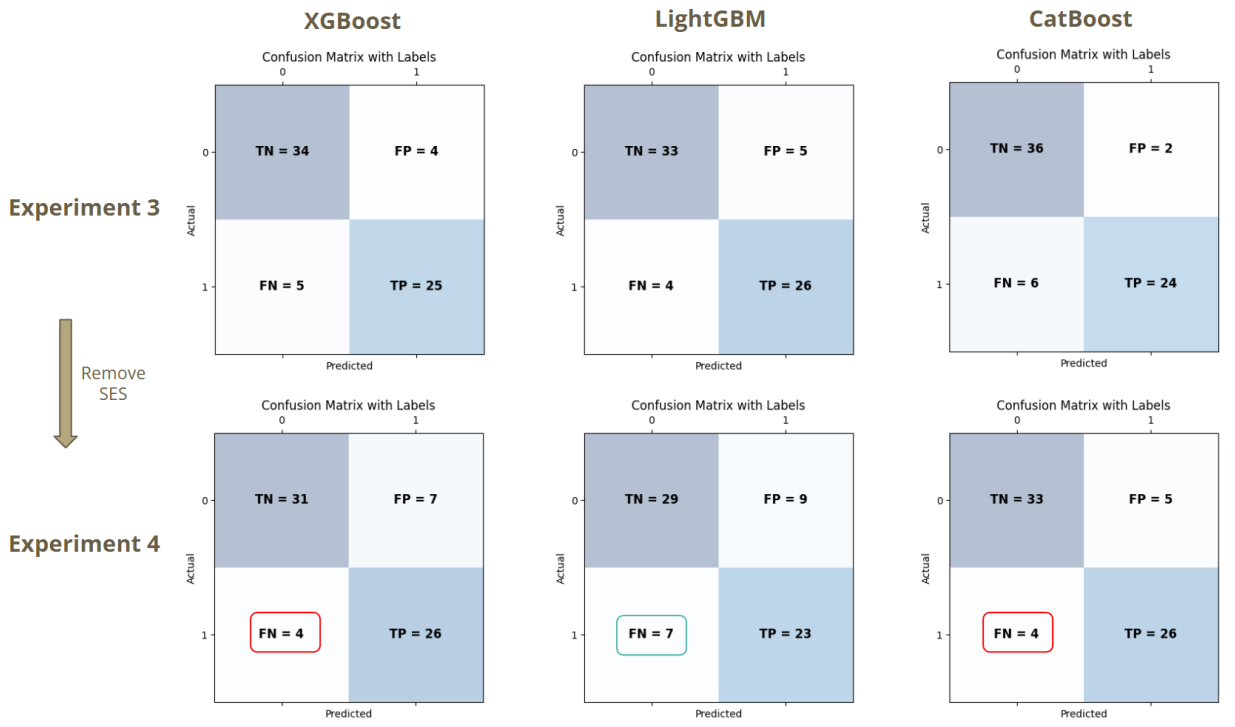


Fig. 23. Experiment 3 vs Experiment 4 Confusion Matrix across the three boosting algorithms

V. Conclusion

In this project, we investigated Alzheimer's classification using tabular data and interpretable machine learning, focusing on both predictive performance and transparency. We first replicated the original study and then extended it by applying fairer evaluation strategies that accounted for potential sources of target leakage.

A critical finding was that commonly used features in most research papers such as CDR and MMSE, while highly predictive, led to artificially inflated performance and thus had to be excluded. After removing these problematic and redundant features, the models still achieved strong results, with accuracies ranging from 86% to 88%, recall from 80% to 86%, precision from 84% to 92%, F1 scores from 84% to 86%, and AUC values from 91% to 96%. Among the three gradient boosting methods evaluated, CatBoost provided the best balance of performance and consistency, achieving 88.24% accuracy, 92.31% precision, 80% recall, 86.71% F1, and 96.05% AUC. Interpretability analysis with SHAP and LIME further revealed that normalized whole brain volume (nWBV), estimated total intracranial volume (eTIV), and years of education (EDUC) were consistently influential predictors, while socioeconomic status (SES) contributed the least. These interpretability tools also helped identify systematic false negatives across models, demonstrating the value of explainable AI in healthcare applications.

The main limitation of this study lies in the relatively small sample size and the limited number of features in the dataset, which constrain the complexity of the models and may limit generalizability.

Future research should address these issues by incorporating additional data modalities such as MRI, PET imaging, or genetic biomarkers to capture the disease's multifaceted nature, as well as validating models on larger and more diverse cohorts like the ADNI dataset. Extending the analysis to longitudinal data could also enable distinguishing stable mild cognitive impairment (MCI) from progressive MCI, thereby supporting earlier and more accurate intervention strategies. Finally, systematic fairness and bias assessments across subgroups such as gender and race remain crucial to ensure that predictive models do not inadvertently perpetuate diagnostic disparities.

Overall, this project highlights that carefully removing biased or misleading features improves model fairness and reliability, and that interpretable machine learning offers both diagnostic accuracy and essential transparency for clinical decision support.

References

[1] "Alzheimer's Disease - UCI MIND." Accessed: Aug. 22, 2025. [Online]. Available: https://mind.uci.edu/dementia/alzheimers/

[2] "Alzheimer's Disease Fact Sheet," National Institute on Aging. Accessed: Aug. 22, 2025. [Online]. Available: https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet

[3] F. Kwan, J. S. Sulistyawan, K. S. Nugroho, and B. Pardamean, "Alzheimer Features for Analysis: An Explainable Gradient Boosting Approach," in *2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, Namakkal, India: IEEE, Apr. 2025, pp. 1–7. doi: 10.1109/AIMLA63829.2025.11041645.

[4] M. Cabanillas-Carbonell and J. Zapata-Paulini, "Evaluation of machine learning models for the prediction of Alzheimer's: In search of the best performance," *Brain, Behavior, & Immunity - Health*, vol. 47, p. 100957, Mar. 2025, doi: 10.1016/j.bbih.2025.100957.

[5] M. G, "Alzheimer Disease Forecasting using Machine Learning Algorithm," *Biosci. Biotech. Res. Comm*, vol. 13, no. 11, pp. 15–19, Dec. 2020, doi: 10.21786/bbrc/13.11/4.

[6] M. Bari Antor *et al.*, "A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–12, July 2021, doi: 10.1155/2021/9917919.

[7] N. Mahendran, P. M. Durai Raj Vincent, K. Srinivasan, V. Sharma, and D. K. Jayakody, "Realizing a Stacking Generalization Model to Improve the Prediction Accuracy of Major Depressive Disorder in Adults," *IEEE Access*, vol. 8, pp. 49509–49522, 2020, doi: 10.1109/ACCESS.2020.2977887.

[8] L. K. Leong and A. A. Abdullah, "Prediction of Alzheimer's disease (AD) Using Machine Learning Techniques with Boruta Algorithm as Feature Selection Method," *J. Phys.: Conf. Ser.*, vol. 1372, no. 1, p. 012065, Nov. 2019, doi: 10.1088/1742-6596/1372/1/012065.

[9] S. Dhakal, S. Azam, K. Md. Hasib, A. Karim, M. Jonkman, and A. S. M. F. A. Haque, "Dementia Prediction Using Machine Learning," *Procedia Computer Science*, vol. 219, pp. 1297–1308, 2023, doi: 10.1016/j.procs.2023.01.414.

[10] R. L. Buckner *et al.*, "A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume," *NeuroImage*, vol. 23, no. 2, pp. 724–738, Oct. 2004, doi: 10.1016/j.neuroimage.2004.06.018.

[11] M. Rosselli, I. V. Uribe, E. Ahne, and L. Shihadeh, "Culture, Ethnicity, and Level of Education in Alzheimer's Disease," *Neurotherapeutics*, vol. 19, no. 1, pp. 26–54, Jan. 2022, doi: 10.1007/s13311-022-01193-z.

[12] Y. Liu, Y. Tan, Z. Zhang, M. Yi, L. Zhu, and W. Peng, "The interaction between ageing and Alzheimer's disease: insights from the hallmarks of ageing," *Transl Neurodegener*, vol. 13, no. 1, p. 7, Jan. 2024, doi: 10.1186/s40035-024-00397-x.

[13] C. R. Beam, C. Kaneshiro, J. Y. Jang, C. A. Reynolds, N. L. Pedersen, and M. Gatz, “Differences Between Women and Men in Incidence Rates of Dementia and Alzheimer’s Disease,” *Journal of Alzheimer’s Disease*, vol. 64, no. 4, pp. 1077–1083, July 2018, doi: 10.3233/JAD-180141.

[14] M. Rosende-Roca *et al.*, “Exploring sex differences in Alzheimer’s disease: a comprehensive analysis of a large patient cohort from a memory unit,” *Alz Res Therapy*, vol. 17, no. 1, p. 27, Jan. 2025, doi: 10.1186/s13195-024-01656-9.

[15] E. J. Canales-Rodríguez *et al.*, “Age- and gender-related differences in brain tissue microstructure revealed by multi-component T2 relaxometry,” *Neurobiology of Aging*, vol. 106, pp. 68–79, Oct. 2021, doi: 10.1016/j.neurobiolaging.2021.06.002.