AUIQ TECHNICAL ENGINEERING SCIENCE 2026;3:109–127

ORIGINAL STUDY

Machine Learning–Driven Prediction of Dementia from MRI and Clinical Features: A Comparative Analysis of Ensemble and Baseline Models

Sarah Raad Hameed *, Zainab Muhannad Nahid, Rawan Ahmed Abdulmahdi

Hilla, Iraq

ABSTRACT

Early detection of dementia remains a pressing challenge in clinical neuroscience, as delayed diagnosis limits therapeutic impact and healthcare planning. Leveraging the Open Access Series of Imaging Studies (OASIS) cross-sectional dataset of 436 participants, this study developed a robust machine learning pipeline integrating sociodemographic, clinical, and neuroimaging-derived features. Preprocessing included removal of highly sparse variables (Delay), median imputation of partially missing but clinically essential measures (SES, MMSE, CDR, Educ), Winsorization of extreme values, and skewness correction. The target Clinical Dementia Rating (CDR) was binarized (0 = no dementia, ≥0.5 = dementia) to align with clinically actionable screening. Categorical features were numerically encoded, and *StandardScaler()* was selectively applied to models sensitive to feature magnitude (Logistic Regression, KNN), while ensemble methods (Random Forest, XGBoost, LightGBM) were trained on raw inputs. Stratified shuffle splits of 70:30 and 60:40 were repeated 20 times, with Synthetic Minority Oversampling Technique (SMOTE) applied exclusively to training sets. Performance was assessed using accuracy, precision, recall, F1-score, specificity, and ROC-AUC, complemented by confusion matrices and ROC curves for the best-performing runs. Results indicated superior performance of ensemble learners, with LightGBM yielding the most balanced outcomes (ROC-AUC = 0.954 ± 0.015; accuracy = 0.890 ± 0.022), while XGBoost achieved the highest recall (0.911 ± 0.026), reducing false negatives. These findings demonstrate that gradient-boosting ensembles provide strong and stable performance under repeated internal validation, supporting their use as a comparative methodological baseline for dementia prediction within the studied cohort.

Keywords: Dementia, Alzheimer's disease, OASIS dataset, Machine learning, LightGBM

1. Introduction

Dementia refers to progressive neurodegenerative syndromes, which compromise cognitive function and overall functioning. About 57 million people were estimated to be living with dementia in 2019, a number that is projected to increase to about 153 million by 2050 primarily due to the effects of population aging [1]. A situation like this provides sufficient basis for establishing a robust rationale involving early risk stratification being an urgent clinical and public health imperative. Performing only marginally superior than age alone, traditional multifactorial risk scores calibrated with a small number of covariates constrains prediction of 10-to-20-year risk due to low detection rates at accessible false-positive thresholds because of disease heterogeneity, multimorbidity, and necessary longitudinal signals [2].

Machine learning (ML) offers nonlinear modeling, temporal representation, and multimodal fusion as promising opportunities. Recent EHR-based transformers encoding event sequences have

Received 18 November 2025; revised 23 January 2026; accepted 6 February 2026.
Available online 26 February 2026

* Corresponding author.
E-mail addresses: Sarahraad200@gmail.com (S. R. Hameed), zainab.muhannad924@gmail.com (Z. M. Nahid), ralwash96@gmail.com (R. A. Abdulmahdi).

https://doi.org/10.70645/3078-3437.1058

demonstrated gains in discrimination with subgroup-aware explanations explicitly auditing fairness across demographic categories [3]. Generalization is further enabled by integrative frameworks combining EHR features with biomedical knowledge networks that embed clinical context [4]. In imaging, multimodal deep neural networks synthesizing MRI and PET for staging disease and predicting its progression achieve competitive discrimination in cohorts such as those found in the ADNI, thus demonstrating the benefit of feature-level fusion [5].

Additionally, time-to-event machine learning models (e.g., DeepSurv) have used population biobanks to predict dementia such that incident modeling can include censoring and thus provide outputs over clinically meaningful time horizons [6, 7]. However, translation to clinical practice is hampered by several known recurrent limitations: variable label validity in routine data; reduced transportability in external cohorts due to calibration drift and dataset shifts across different care settings; biases and inequities, risk of data leakage, and ultimately the lack of robust decision-analytic evidence on net benefit [8–13]. Leading studies addressing these challenges through transparent pipelines implemented under the TRIPOD-AI framework with prespecified external validation including recalibration, explainable models (e.g., attribution analyses), and decision-curve analyses connecting predictions to actionable outcomes [13–16].

The central objective of this article is to design and evaluate a robust machine learning pipeline for the early detection of dementia using sociodemographic, clinical, and neuroimaging-derived features from the OASIS dataset. A key methodological focus was addressing class imbalance through the use of the Synthetic Minority Over-sampling Technique (SMOTE), applied exclusively to the training folds, thereby enhancing learning while preserving unbiased validation. In addition to comprehensive preprocessing—comprising median imputation for missing values, winsorization to mitigate outlier effects, skewness transformation, categorical encoding, and selective feature standardization via StandardScaler()—the framework was optimized to reduce diagnostic errors. Particular emphasis was placed on minimizing false negatives, as the clinical cost of missed dementia cases is substantially greater than that of false alarms. By integrating these strategies, the framework aims to establish a reproducible and generalizable predictive tool that not only improves overall model performance but also prioritizes sensitivity to ensure timely clinical recognition and intervention in individuals at risk.

2. Literature review

The use of ML and DL for dementia prediction has advanced into a highly methodological, cross-modal, interpretable, and inclusive approach. Foundational proof-of-concept studies indicated that computational models could be populated with well-accepted neuropsychiatric and neuroimaging markers to predict cognitive decline. One of the recent works Gill et al. (2020) [17] demonstrates well over 80% accuracy in prediction when minor behavioral impairment domains are included along with hippocampal volume. The finding underscores both the early signaling capacity of sustained neuropsychiatric symptomatology and the pathology of aging. This impactful approach was, however, limited more by relatively small but imaging-rich cohorts and hence further underscores the imperative need for approaches that can be taken to scale in larger populations.

Multi-layer perceptron and convolutional recurrent architectures were applied on ADNI data revealing that feedforward deep learning architectures (multi-layer perception) are the most effective in classification of normal cognition, mild cognitive impairment (MCI), and dementia. Multi-layer perception was compared by Stamate et al. (2020) [18]. However, more complex recurrent architectures have been revealed to be vulnerable to sparse longitudinal data, thus posing an even larger question on how temporal information can be leveraged from clinical datasets that are irregular or incomplete.

The studies that took advantage of the national health databases further brought progress closer to real-world implementation. Kim et al. (2021) [19] used a deep neural network on more than 160,000 records from the Korean National Health Insurance cohort and outperformed the accuracies of conventional ML classifiers thereby establishing that DL could be applied efficiently at the scale of administrative data. However, scaling up in this way also meant losing clinical granularity while gaining a different kind of statistical power.

In contrast, Guram et al. (2021) [20] showed that classical ensemble models like Random Forests were superior to smaller curated datasets with accuracies above 94%. This creates a methodological paradox: Deep Learning performs optimally with large heterogeneous datasets, but simpler ensemble models outperform it on small well-structured data requiring flexible context-sensitive model selection, not an overarching algorithmic hierarchy.

More integrative frameworks have begun to address this disparity. As Sakatani et al. (2022) [21] reviewed, in other words, incorporating various

modalities—MRI, spectroscopy, and some basic blood markers—consistently demonstrated that ML models achieved or surpassed clinician-level accuracy. Deep neural networks described by Oyama and Sakatani trained on time-resolved near-infrared spectroscopy and blood tests offered feasibility demonstration low-cost mass-deployable screening. Progress like this only highlights into even sharper relief the structural gap: advanced performance continuously comes at the cost of interpretability, hence clinical trust.

High discrimination is attainable with transparency regarding feature contributions, as usable frameworks based on LightGBM and SHAP values introduced by You et al. (2022) [22] testified. Li et al. (2024) [23] take further the interpretability advances by inserting decision-focused selection layers inside deep models over unstructured clinical notes directly speaking to the issue of black box. These innovations mark a fundamental change in the field: model explainability is ever more often considered a precondition for clinical translation rather than an add-on.

At the same time, there has been conceptual broadening of valid predictors of dementia. As demonstrated in the works of Zadgaonkar et al. (2023) [24] and Song et al. (2024) [25], psychosocial and emotional variables—physical activities, smoking behaviors, depression, and loneliness—achieve predictive accuracies equal to those obtained through biomarkers, except that most earlier models were framed biomedically. This is indicative of the fact that vulnerability to dementia is deeply embedded within modifiable behavioral and social pathways; therefore, reframing offers substantial practical relevance when new models can identify measurable yet previously ignored predictors that are also actionable and thereby join prediction with prevention through community-based interventions or public health strategies. However, such models would need validation using longitudinal data to measure actual reduction in future dementia risk plus biomedical data for robustness since current analyses remain largely cross-sectional and self-reported.

The field has recently advanced toward inclusivity and equity. This has been a long-standing weakness of the dementia research agenda, such that the majority of studies focused on predominantly White, high-income populations. Gradient boosting models trained by Ports et al. (2025) [26] using American Indian and Alaska Native cohorts provided AUROCs greater than 0.9 while tuning predictions to the epidemiological realities of marginalized groups. Similarly, Akter et al. (2025) [27] used more than 230,000 electronic health records from OneFlorida + to predict Alzheimer's disease and related dementias two years before diagnosis thereby demonstrating both scalability and clinical utility in highly diverse real-world contexts. The efforts amount to a critically needed interventional pivot embedding AI-driven dementia prediction within precision health frameworks oriented around previously underrepresented populations. Unfortunately, the broad literature remains imbalanced with interpretability-awareness as well as equity-awareness pursued mostly in isolation rather than systematically integrated into coherent frameworks.

Methodologically, the field has moved from models that throughout most of its existence were constrained by features to sophisticated multimodal frameworks that would be interpretable and able to handle large volumes of data. Conceptually, it evolved from an understanding based in biomarkers to more comprehensive models that incorporate lifestyle factors, psychosocial elements, and social determinants of health. Theoretically, efforts at structural reform are directed toward the reduction of population bias through the development of culturally and demographically predictive tools. Methodological gaps remain significant, nonetheless. Longitudinal modeling is hindered not just by irregular follow-up but also by data sparsity. Interpretability is typically added as a retrospective consideration rather than being built into model architecture from inception despite recent marginal improvements. Further hybridization of biomedical- and genetic-based markers with high-potential lifestyle-based predictors takes place within the framework at a level still primarily conceptual. This places the discipline at a very pivotal point: further studies should focus on algorithm enhancements integrated into infrastructures that are clinically actionable, socially equitable, and ethically transparent. Only through the integration of such aspects into predictive modeling can it attain its potential as a transformative force in the prevention and care of dementia.

3. Methodology

This study followed a structured and sequential approach intended to make the study reproducible, transparent, and clinically interpretable. The steps of the process include loading the data set and initial exploration through which descriptive statistics are generated to establish baseline distributions and identify anomalies. This is accompanied by a focused exploratory data analysis (EDA) that would assist in visualizing the properties of the population over which dimensions such as demographic, clinical, and neuroimaging variables have been measured so as to make informed preprocessing decisions.

Table 1. Dataset features with understandable names and explanations.

Original Feature	New Understandable Name	Short Explanation
ID	Participant ID	Unique identifier assigned to each subject in the study to ensure anonymization and tracking.
M/F	Biological Sex	Denotes whether the participant is male (M) or female (F), an important demographic variable influencing dementia risk.
Hand	Dominant Hand	Records whether the subject is right-handed or left-handed, sometimes linked to brain lateralization and cognitive processing.
Age	Age (Years)	The participant's age in completed years; increasing age is the strongest risk factor for dementia.
Educ	Education Level (Years)	Number of years of formal education completed, used as a proxy for cognitive reserve and resilience against neurodegeneration.
SES	Socioeconomic Status (Scale)	Socioeconomic standing, often scored (e.g., 1–5), reflecting access to resources, healthcare, and lifestyle factors impacting brain health.
MMSE	Mini-Mental State Examination Score	A 30-point standardized cognitive screening test; lower scores indicate greater cognitive impairment.
CDR	Clinical Dementia Rating (Target Variable)	A clinician-rated global dementia severity scale (0 = none, 0.5 = very mild, 1 = mild, 2 = moderate, 3 = severe). Commonly used as the ground-truth label.
eTIV	Estimated Total Intracranial Volume (mL)	MRI-derived measure of total cranial capacity, serving as a normalization factor for brain size.
nWBV	Normalized Whole Brain Volume (%)	Ratio of whole brain tissue volume to intracranial volume, expressed as a percentage; lower values indicate brain atrophy.

Preprocessing was executed sequentially. The highly sparse Delay feature was dropped and variables SES, MMSE, CDR, and Education imputed by their medians. Duplicate records were checked and removed if found. Outliers were treated by Winsorizing at a particular level so as not to mar the real distribution and create substantial distortion. Skewness was found and treated accordingly with transformation techniques. Categorical sex and handedness variables have been encoded in binary form. Selective feature scaling has been applied using StandardScaler() only for those models which are sensitive to the magnitude of features. After data refinement, a modeling pipeline was built using 70:30 and 60:40 stratified splits repeated over 20 runs each. The Synthetic Minority Oversampling (SMOTE) was used only on the training sets to avoid leakage, in resolving the class imbalance problem. Further comparative evaluation and tuning of algorithms — baseline models comprising Logistic Regression and K-Nearest Neighbors (KNN) alongside more advanced ensemble methods such as Random Forest, XGBoost, and LightGBM — was carried out. A comprehensive assessment plan was put into action, reporting metrics like accuracy, precision, recall, F1-score, specificity, and ROC-AUC, backed by diagnostic visuals such as confusion matrices and ROC curves. Every one of these steps is explained in detail in the next sections of the methodology

3.1. Dataset loading and initial exploration

The data used in this study was obtained from the Open Access Series of Imaging Studies (OASIS), a public resource allowing studies on dementia and neurodegeneration [28]. The analytic cohort contains 436 subjects with and without dementia which creates a balanced foundation for comparison analysis. Only th cross-sectional part is used here so that a concise representation of the variation, clinical, and neuroimaging can be obtained while avoiding the added complication of long-term follow-up. It comprises sociodemographic attributes by sex, age, education, socioeconomic status, clinical measures (Mini-Mental State Examination and Clinical Dementia Rating), and neuroimaging derived features (two estimated total intracranial volumes; normalized whole brain volume; atlas scaling factor). In this presentation, essential predictors for the classification of dementia that have also been reformulated in more interpretable terms to cater clarity aspect both domains—the clinical domain and the computational domain—will be used. The mapping of all original features with their understandable names and explanations is presented fully in Table 1 – Dataset Features with Understandable Names and Explanations.

3.2. Exploratory data analysis (EDA)

Five figures most effectively represent distinct yet complementary aspects of the dataset. Figs. 1 and 2, Distribution of Dementia Outcomes in the Study Cohort and Gender Distribution of Participants present two pie charts describing the composition of cases and the demographic balance within the sample. The current case-enriched sample comprises 301 dementia cases (69%) and 135 controls (31%), with more men than women, 268 men (61.5%) to 168

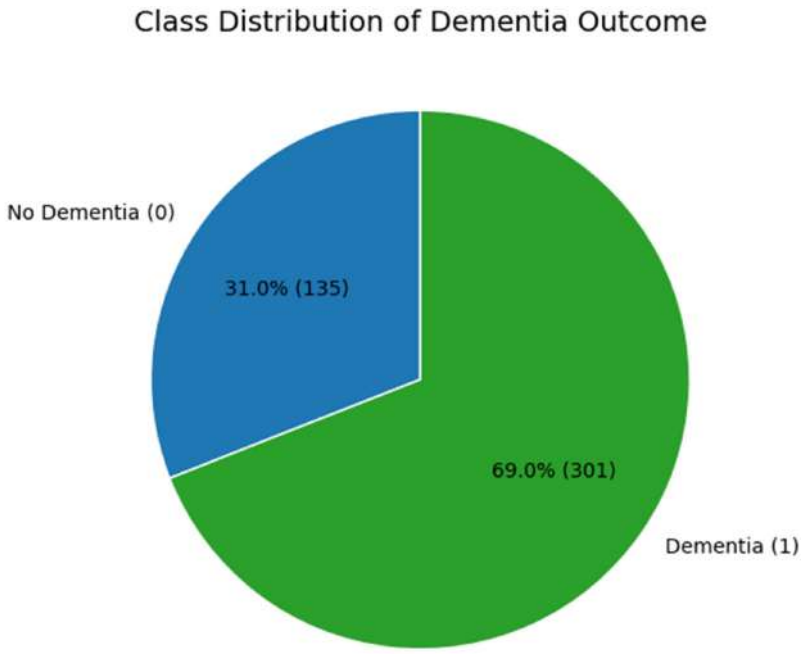

Fig. 1. Distribution of dementia outcomes in the study cohort.

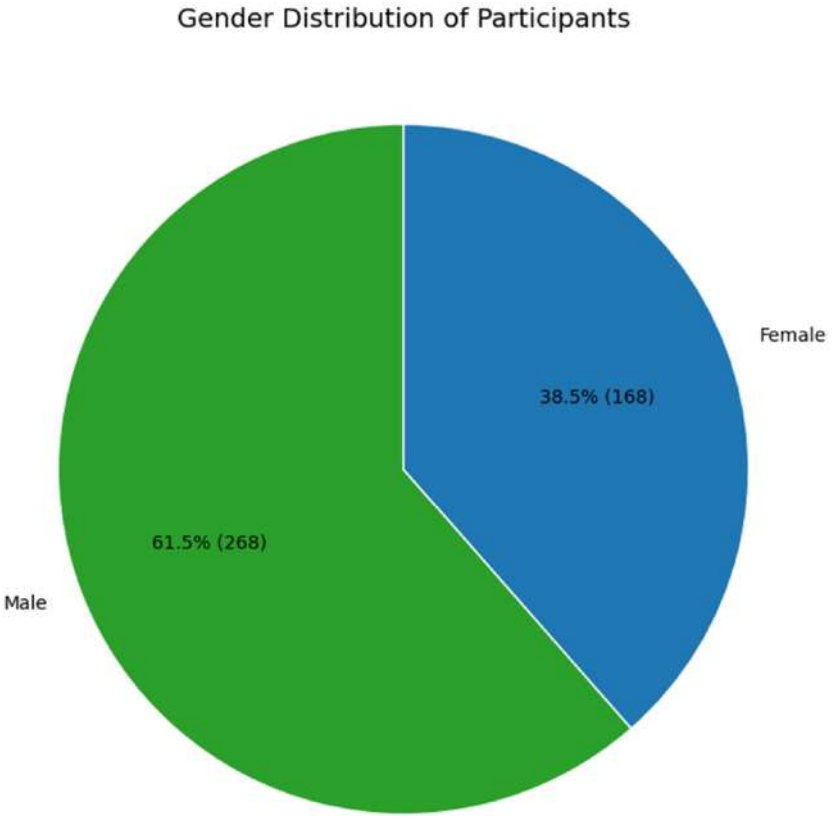

Fig. 2. Gender distribution of participants.

women (38.5%). Such visualizations bring to fore very critical imbalances that caution left aside can insidiously propagate into predictive models through mechanisms such as weighting or covariate adjustment rather than explicit stratification. Fig. 3 presents correlation pairs among six numerical features. The two major relationships: firstly, a near-perfect negative correlation between eTIV and ASF ($r \approx -0.98$)

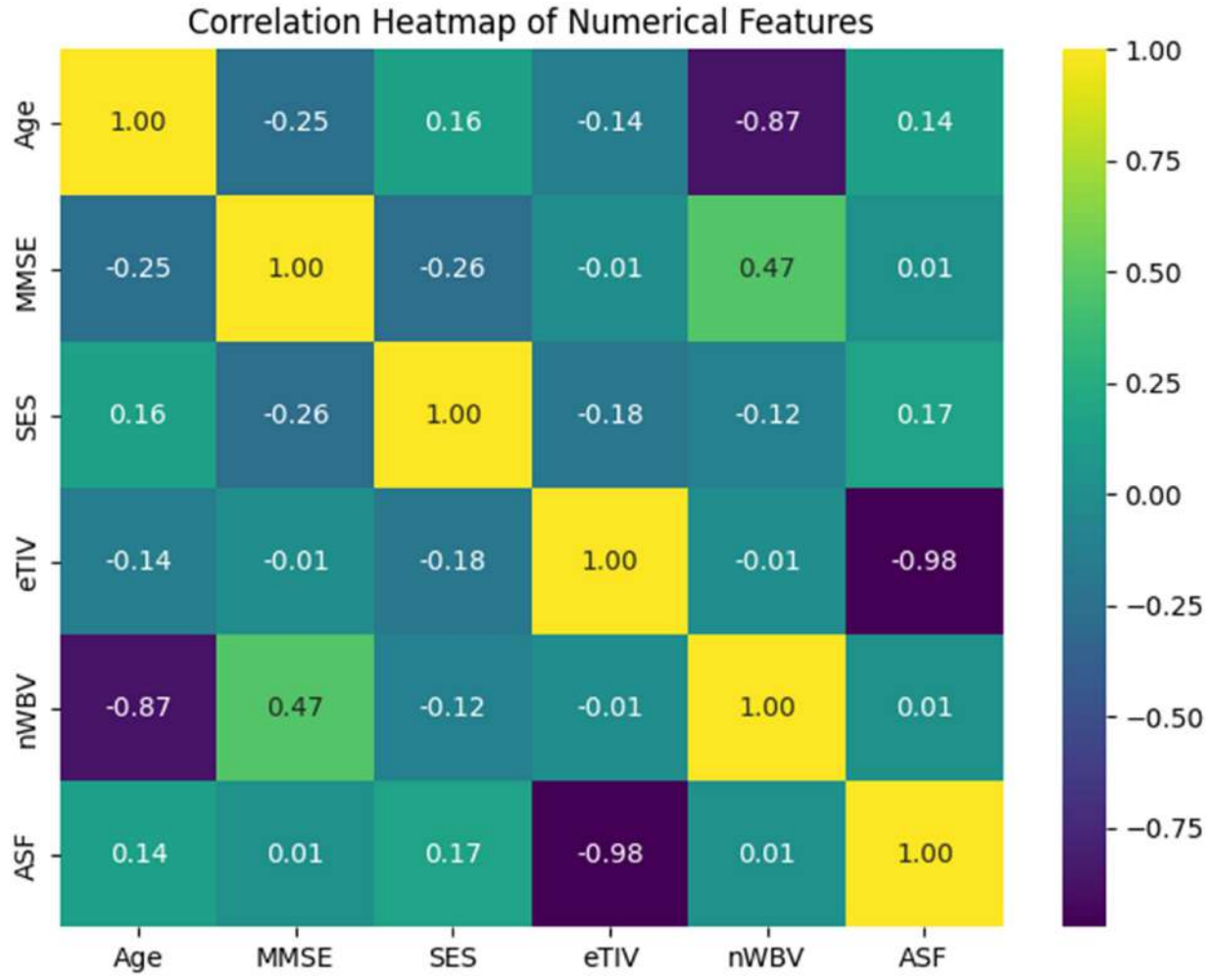

Fig. 3. Correlation matrix of clinical and neuroimaging variables.

demonstrate redundancy in metrics of head-size, and secondly, a strong negative correlation between Age and nWBV ($r \approx -0.87$), validates the premise that brain volume decreases with age as affirmed by the elderly participants in this study. Also, MMSE shows a moderate positive correlation with nWBV ($r \approx +0.47$). The others are weaker negatives with Age and SES. All other pairwise correlations are very small to indicate a concern of multicollinearity but support selective feature inclusion. Fig. 4 is a clear demonstration that there are non-normal, multimodal distributions since the bimodality of Age is quite evident about ceiling effects for MMSE and mixture-like patterns for nWBV. Fig. 5 displays boxplots for six features by binary outcomes, with MMSE showing the greatest separation—lower median scores for cases indicating dementia— but age and nWBV presenting unexpected shifts in direction, thus requiring verification of coding and adjustments for confounders. In sum, this set of figures effectively articulates those fundamental aspects of the cohort: class imbalance; gender skew; and multicollinearity between measures of head size together with non-Gaussian distributions of features that must form critical elements of context against which to judge both the modeling analyses themselves and their robustness.

3.3. Data preprocessing

3.3.1. Handling missing values

A dataset completeness audit was conducted as part of the data integrity assessment and is summarized in Table 2 (Dataset Completeness and Missing Value Summary by Feature). The missingness pattern was heterogeneous: several variables were fully complete (e.g., Age, eTIV, nWBV, ASF, and categorical demographics), whereas certain clinically relevant attributes exhibited substantial incompleteness. The feature *Delay* was severely sparse, with >95% missing values, and was therefore removed to avoid unreliable inference and instability in downstream modeling.

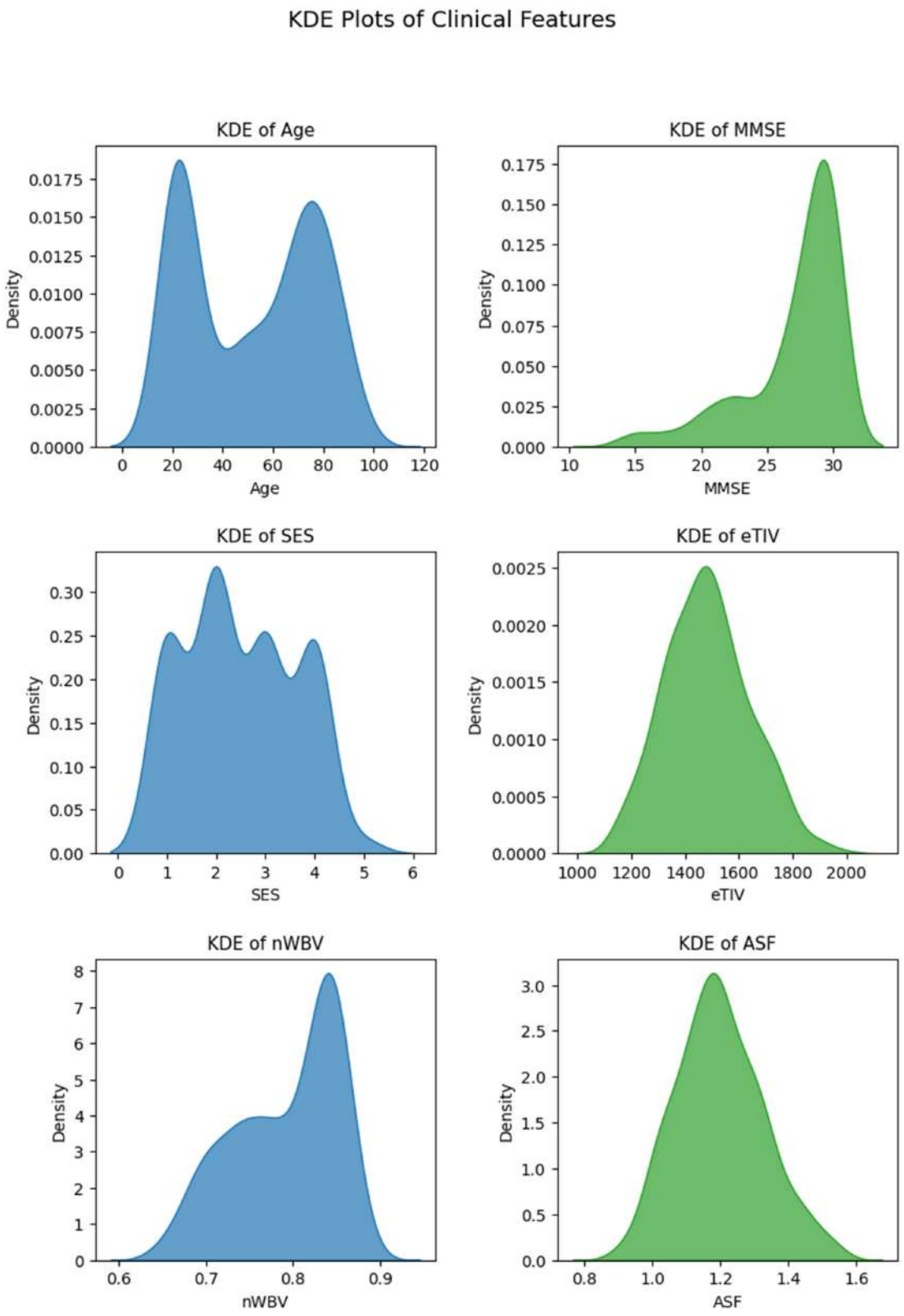

Fig. 4. Kernel density estimates of age, cognitive, and neurostructural features.

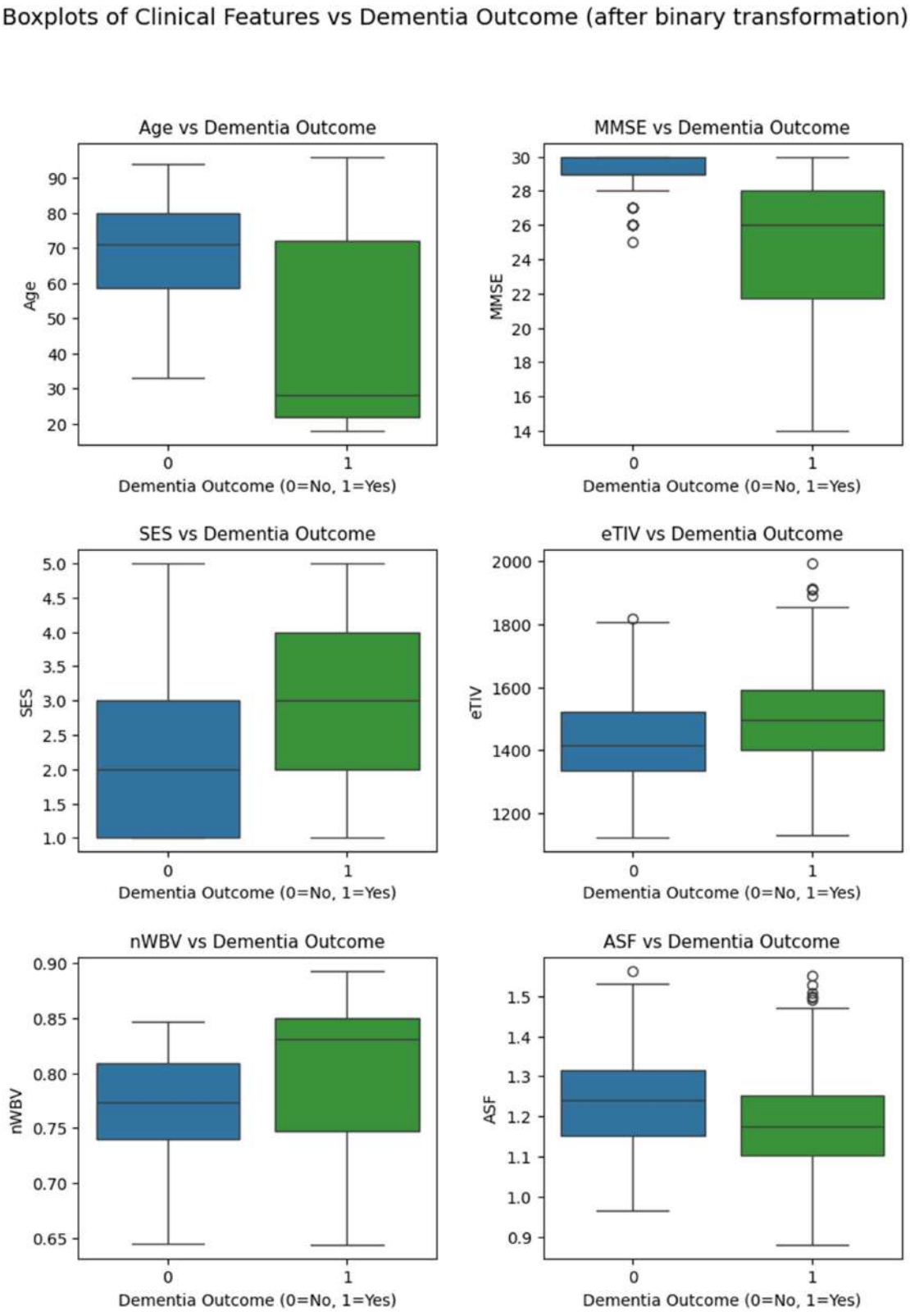

Fig. 5. Comparative distributions of clinical features by dementia status.

Table 2. Dataset completeness and missing value summary by feature.

Feature	Data Type	Non-Null Count	Missing Values (Count)	Missing Values (%)	Unique Values
ID	Object	436	0	0.00	436
M/F	Object	436	0	0.00	2
Hand	Object	436	0	0.00	1
Age	Integer	436	0	0.00	73
Educ	Float	235	201	46.10	5
SES	Float	216	220	50.46	5
MMSE	Float	235	201	46.10	17
CDR	Float	235	0	46.10	4
eTIV	Integer	436	0	0.00	312
nWBV	Float	436	0	0.00	182
ASF	Float	436	0	0.00	282
Delay	Float	20	416	95.41	14

Note: Because CDR is the ground-truth label, records with missing CDR were excluded prior to preprocessing. Therefore, subsequent preprocessing and modeling were performed on N = 235 records, and CDR was not imputed.

Target-label integrity was treated as a strict constraint. Because the Clinical Dementia Rating (CDR) constitutes the ground-truth outcome in this study, CDR was excluded from all imputation procedures. Instead, all records with missing CDR values were removed prior to any preprocessing or model development, preventing label leakage and preserving the validity of supervised learning. After this exclusion step, missing values in predictor variables—specifically SES, MMSE, and Educ—were imputed using the median, a robust choice commonly adopted in biomedical datasets where distributional normality cannot be assumed and outliers may be present.

Overall, this strategy—(i) removing *Delay* due to extreme sparsity, (ii) excluding missing-target records for CDR, and (iii) applying robust median imputation to remaining incomplete predictors—balances methodological rigor with maximal retention of clinically informative signals.

3.3.2. Duplicate record detection and integrity verification

A formal duplicate detection process was carried out as the first step of pre-processing to maintain data integrity and remove redundancy. All rows were fully compared so that any potential repeated observation that might lead to biased distributional properties, inflated sample size, or violated the validity of later statistical inference and model training could be detected. Specifically, a row-wise duplicate assessment was implemented using Python's drop_duplicates() function by which feature values for all records are compared. The output in summary indicated 436 records in the dataset before and after dropping duplicates; hence, no duplicate entries were present. Thus, the dataset has already been established to be nonredundant and therefore preserves the integrity of further analysis with not required for more corrective measures.

3.3.3. Outlier detection and treatment

Outlier management was conducted conservatively to maintain clinically meaningful variability as well as the robustness of analyses. A first review using the standard IQR-based approach flagged almost half the observations as outliers, thereby reducing the dataset from 436 records to 211. While this method has gained popularity in statistical preprocessing, such a considerable loss of data cannot be accommodated; hence, it is considered highly restrictive and can compromise the representativeness of a cohort. The problem here henceforth be addressed by adopting an even more conservative treatment through the Winsorization of extreme values at both ends of the distribution rather than elimination. This adjustment preserved the total number of records (436 before and after treatment) but reduced the undue effect that extreme observations were having on the results. For that reason, henceforth, Winsorization was adopted as the final strategy since it had proven to be an effective compromise approach in an attempt both to reduce the effects of outliers and not to spectrum all clinically relevant information.

3.3.4. Skewness analysis

Another step of preprocessing included a review of the distributional symmetry of numeric variables since skewness could greatly impact statistical inference and model results. Results are summarized in Table 3 (Skewness Analysis of Numerical Features with Threshold-Based Interpretation) where it is demonstrated how each feature was classified as approximately symmetric, moderately skewed, or highly skewed and prior transformation indicated that several features were suboptimal asymmetric:

Table 3. Skewness analysis of numerical features with threshold-based interpretation.

Feature	Skewness Value	Interpretation
Age	−0.0047	Approximately Symmetric – No Action Needed
Educ	0.2364	Approximately Symmetric – No Action Needed
SES	0.9211	Moderate Skew – Transformation Optional
MMSE	−2.5947	High Skew – Transformation Recommended/Applied
CDR	1.8381	High Skew – Transformation Recommended/Applied
eTIV	0.2224	Approximately Symmetric – No Action Needed
nWBV	−0.5093	Moderate Skew – Transformation Optional
ASF	0.2830	Approximately Symmetric – No Action Needed

the Mini-Mental State Examination (MMSE) is revealed to be strongly negatively skewed while the Clinical Dementia Rating (CDR) is revealed to be highly positively skewed, and thus adjustment corrective action is necessitated. Thus, features that exceeded the high-skew threshold were selectively subjected to automatic transformation techniques so that their distributions would more closely resemble symmetry. This brought the skewness of MMSE up from −2.59 to −0.72, putting it in the "moderate skew" category and reduced CDR's skewness from 1.84 to 1.70, less extreme but positive nonetheless. Other variables included Age, Education (Educ), Socioeconomic Status (SES), estimated Total Intracranial Volume (eTIV), normalized Whole Brain Volume (nWBV), and Atlas Scaling Factor (ASF) had either approximate symmetry or moderate skew-which does not make transformation mandatory. This also keeps the data straightforward to interpret while making it more appropriately suited for later modeling steps by lessening the overly strong influence that highly skewed features can have.

3.3.5. Target transformation of clinical dementia rating (CDR)

The CDR has been recoded as a binary outcome variable from the original multiclass scale for purposes of predictive modeling. There were four gradations in the original CDR classification: 0 = no dementia, 0.5 = very mild Alzheimer's disease, 1 = mild Alzheimer's disease, and 2 = moderate Alzheimer's disease. Though this fine scale has tremendous value in offering insight clinically, direct implementation in machine learning leads to two major challenges: (i) high class imbalance especially among the higher severity categories and (ii) need for simplification of the prediction into a clinically actionable framework. Therefore, the outcome too was recoded into two categories to address both considerations: 0 = no dementia (CDR = 0) and 1 = dementia (CDR $\geq$ 0.5). This binarization does not only mitigate the sparsity problem that is associated with minority classes but also maps the target variable to the main clinical decision-making problem, that is, whether dementia exists or not. In this way, it increases both statistical robustness and translational relevance of early detection models.

3.3.6. Encoding of categorical features

To meet the numeric requirement input of machine learning algorithms, categorical variables were systematically recoded into binary codes. Sex was coded as male = 1, female = 0; therefore, a standard numeric code for recording the sex of participants was created. Similarly, the *Hand* variable was recorded only as right- or left-handed and subsequently encoded with right = 1 and left = 0. This preserves the semantics of the original attributes appropriate for training the model. However, it should be noted that there exists only a single category for handedness in this dataset (all participants are right-handed); hence, this feature is unlikely to contribute analytically in terms of discrimination and thus ultimately omitted from all subsequent modeling steps. The test ensured consistent data encoding by removing non-numeric restrictions while still maintaining clinically meaningful demographic distinctions.

3.3.7. Standardization of model inputs

To ensure comparable feature scales and reduce the dominating effect made by those variables which have larger numeric ranges, standardization has been applied selectively with the application of StandardScaler (). This transformation rescales each feature to have zero mean and unit variance, thus making optimization stable as well as distance computations meaningful. StandardScaler () was applied to those models that are inherently sensitive to magnitudes of features i.e. Logistic Regression and K-Nearest Neighbors (KNN). The raw features were supplied into the tree-based ensemble and boosting methods, i.e. Random Forest, XGBoost, and LightGBM since their splitting mechanism is not affected by any monotonic transformation of input scales. This deliberate use of StandardScaler () ensured methodological rigor by matching the preprocessing with the mathematical requirements of the algorithms and hence improving convergence stability as well as fairness in performance evaluation across models.

3.3.8. Comprehensive modeling pipeline and performance evaluation strategy

A hybrid modeling workflow and performance evaluation approach were adopted in such a way that led to systematic equilibrium between data preprocessing, model training, and comparison. The data set was split with stratified shuffle splits of test proportions 70:30 and 60:40 — each repeated across 20 independent runs — for the purpose of reducing partition-induced variance as well as obtaining relatively stable estimates of performance. The Synthetic Minority Over-Sampling Technique (SMOTE) was applied only on training folds so there would never be any possibility of validation leakage and therefore addressing class imbalance. Feature scaling using standardization by z-score was selectively applied to the models that are sensitive to feature magnitudes — for instance, logistic regression, and k-nearest neighbors, while tree-based methods like random forest, XGBoost, and LightGBM shall be trained based on distribution over raw features since they are invariant to scale. For Hyperparameter settings: Models were trained using default configurations, with only basic reproducibility-related settings (e.g., random seeds and convergence limits where applicable), and no exhaustive hyperparameter optimization was performed. For every run, evaluation of predictions was based on a very broad suite of performance metrics which included accuracy, precision, recall (sensitivity), F1-score, specificity, and ROC-AUC. Other indices derived from the confusion matrix such the false-positive rate and false-negative rate were also included. Results were aggregated across runs to capture central tendency as well as variability and reported as mean ± standard deviation. In addition, for facilitation of interpretation and readability in diagnostics, the best-performing run for each model in terms of the highest ROC-AUC will be visualized through heatmaps of confusion matrices and receiver operating characteristic (ROC) curves. For validation scope: All experiments in this study were conducted as internal validation using repeated stratified train–test splits (70:30 and 60:40) on the same dataset. No external cohort or independent dataset was used, and therefore the reported performance reflects internal generalization only. This integrated approach not only facilitates ensure reproducibility and fairness in model comparison but underlines the robustness of ensemble-based classifiers over traditional baselines across varying data partitions; all metrics that are reported here mathematically are introduced in the subsequent section for issues related to transparency and reproducibility.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [29]$$

$$Precision = \frac{TP}{TP + FP} \quad [29]$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \quad [29]$$

$$F1 - score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad [29]$$

$$Specificity = \frac{TN}{TN + FP} \quad [30]$$

$$ROC - AUC = \int_0^1 TPR\,(FPR)\,d\,(FPR) \quad [31]$$

4. Results and discussion

Across 20 repeated stratified runs, the ensemble models from the gradient-boosting family consistently outperformed classical baselines, establishing LightGBM and XGBoost as the dominant classifiers (Table 4). LightGBM achieved the best overall sorting in both split ratios, reaching ROC-AUC scores of 0.954

Table 4. Evaluation metrics for stratified 70:30 and 60:40 splits (20 runs, mean ± std).

Model	Split	Accuracy (mean ± std)	Precision (mean ± std)	Recall (Sensitivity) (mean ± std)	F1-score (mean ± std)	Specificity (mean ± std)	ROC-AUC (mean ± std)
Logistic Regression	70:30	0.855 ± 0.028	0.938 ± 0.022	0.846 ± 0.037	0.889 ± 0.023	0.876 ± 0.046	0.936 ± 0.017
KNN	70:30	0.856 ± 0.028	0.925 ± 0.023	0.860 ± 0.041	0.891 ± 0.023	0.846 ± 0.053	0.917 ± 0.022
Random Forest	70:30	0.883 ± 0.022	0.941 ± 0.023	0.887 ± 0.028	0.912 ± 0.017	0.876 ± 0.052	0.951 ± 0.016
XGBoost	70:30	0.890 ± 0.020	0.931 ± 0.024	0.908 ± 0.025	0.919 ± 0.015	0.850 ± 0.058	0.951 ± 0.014
LightGBM	70:30	0.890 ± 0.022	0.939 ± 0.027	0.899 ± 0.029	0.918 ± 0.017	0.870 ± 0.060	0.954 ± 0.015
Logistic Regression	60:40	0.855 ± 0.025	0.932 ± 0.022	0.853 ± 0.031	0.891 ± 0.020	0.860 ± 0.048	0.934 ± 0.015
KNN	60:40	0.854 ± 0.034	0.925 ± 0.022	0.860 ± 0.046	0.890 ± 0.027	0.842 ± 0.050	0.917 ± 0.021
Random Forest	60:40	0.883 ± 0.024	0.937 ± 0.024	0.892 ± 0.036	0.913 ± 0.019	0.864 ± 0.058	0.950 ± 0.015
XGBoost	60:40	0.888 ± 0.018	0.927 ± 0.022	0.911 ± 0.026	0.918 ± 0.013	0.837 ± 0.055	0.952 ± 0.010
LightGBM	60:40	0.887 ± 0.022	0.927 ± 0.019	0.907 ± 0.029	0.917 ± 0.017	0.840 ± 0.047	0.955 ± 0.011

Table 5. Confusion matrix metrics for stratified 70:30 and 60:40 splits (20 runs, mean ± std).

Model	Split	TN (mean ± std)	FP (mean ± std)	FN (mean ± std)	TP (mean ± std)	FPR (mean ± std)	FNR (mean ± std)
Logistic Regression	70:30	35.900 ± 1.895	5.100 ± 1.895	13.850 ± 3.351	76.150 ± 3.351	0.124 ± 0.046	0.154 ± 0.037
KNN	70:30	34.700 ± 2.170	6.300 ± 2.170	12.600 ± 3.666	77.400 ± 3.666	0.154 ± 0.053	0.140 ± 0.041
Random Forest	70:30	35.900 ± 2.142	5.100 ± 2.142	10.200 ± 2.502	79.800 ± 2.502	0.124 ± 0.052	0.113 ± 0.028
XGBoost	70:30	34.850 ± 2.372	6.150 ± 2.372	8.300 ± 2.283	81.700 ± 2.283	0.150 ± 0.058	0.092 ± 0.025
LightGBM	70:30	35.650 ± 2.455	5.350 ± 2.455	9.050 ± 2.578	80.950 ± 2.578	0.130 ± 0.060	0.101 ± 0.029
Logistic Regression	60:40	46.450 ± 2.578	7.550 ± 2.578	17.750 ± 3.806	103.250 ± 3.806	0.140 ± 0.048	0.147 ± 0.031
KNN	60:40	45.450 ± 2.711	8.550 ± 2.711	17.000 ± 5.568	104.000 ± 5.568	0.158 ± 0.050	0.140 ± 0.046
Random Forest	60:40	46.650 ± 3.135	7.350 ± 3.135	13.100 ± 4.323	107.900 ± 4.323	0.136 ± 0.058	0.108 ± 0.036
XGBoost	60:40	45.200 ± 2.960	8.800 ± 2.960	10.800 ± 3.124	110.200 ± 3.124	0.163 ± 0.055	0.089 ± 0.026
LightGBM	60:40	45.350 ± 2.515	8.650 ± 2.515	11.200 ± 3.473	109.800 ± 3.473	0.160 ± 0.047	0.093 ± 0.029

± 0.015 at the 70:30 split and 0.955 ± 0.011 at the 60:40 split, along with high accuracy (0.890 ± 0.022; 0.887 ± 0.022) and F1-scores (0.918 ± 0.017; 0.917 ± 0.017). This shows that LightGBM has a strong mix of sorting and setting making it the most robust general-purpose learner across runs.

XGBoost showed better performance, getting recall scores of (0.908 ± 0.025) and (0.911 ± 0.026) for these two splits plus even higher F1-scores (0.919 ± 0.015; 0.918 ± 0.013). Greater sensitivity led to a lower number of false negatives on average (FN = 8.3 at 70:30 and 10.8 at 60:40) with specificity being somewhat less (0.850 and 0.837) accompanied by slightly more false positives than LightGBM (Table 5). These trade-offs delineate distinct utility domains: LightGBM emerges as the more balanced default option, whereas XGBoost offers advantages in contexts where minimizing missed detections is paramount.

Random Forest further reinforced the strength of ensemble-based methods, delivering the highest specificity (0.876 at 70:30 and 0.864 at 60:40), a strong tendency toward the limitation of false positives, still at the expense of recall which never overtook boosting algorithms. This is best seen in the optimal runs confusion matrices where boosting models have made fewer false positive decisions than Random Forest but identified more true cases correctly; thus, sensitivity has been better with lower specificity. In addition to being less accurate (as measured with F1-score), Logistic Regression and K-Nearest Neighbors were found to be substantially more variable, having by far the largest spread in accuracy for KNN (std = 0.034 at 60:40 split). Such instability proves the linear and distance-based models to be inferior to tree-based ensemble resilience under changing partition structures.

Quantitative patterns were validated by visualizing the best run performances. In the 70:30 split, XGBoost produced five false negatives (six in the case of LightGBM) with three false positives (two in the case of LightGBM). This demonstrated the predictive strength of both boosting algorithms since they were able to attain 114 true positive values out of only seven false negatives at a split of 60:40. The ROC curve for best runs approached an AUC value of 0.978 which further proves their capability in terms of discrimination between classes and validates the reliability prediction consistency across several runs (Fig. 8a-e and 9a-e). Aggregate comparative bar plots tend to provide more elaboration on this trend where one such bar chart summarizes cross-metric performance at a 60:40 split, and another provides similar results at a 70:30 split, hence confirming that LightGBM is always better as far as AUC and specificity are concerned whereas XGBoost leads slightly in terms recall and F1, Random Forest comes out Maximum Specificity however with much reduced sensitivity.

The results set a distinct ranking. LightGBM gets the best overall trade-off across all metrics, so it becomes the preferred default for general use where false positives are more important. XGBoost can be recommended practically as a first choice in clinical screening applications when missing a positive is substantially more critical than raising alarms because it has slightly better recall (which means it finds more cases). Random Forest is still attractive if one seeks to make decisions very conservatively and keep false positives low. These insights are reinforced by both quantitative outcomes (Tables 4 and 5) and visual evidence (Figs. 6 to 11), collectively affirming the robustness of ensemble methods under repeated internal evaluation in this predictive framework.

It should be noted that the reported results represent internal validation based on repeated stratified splits of the same dataset; external validation was not performed and is beyond the scope of the present study.

While boosting-based learners (e.g., LightGBM/XGBoost) are often considered amenable to interpretability through feature-importance and SHAP-based explanations, the present study focuses on predictive performance and internal validation, and no dedicated explainability analysis

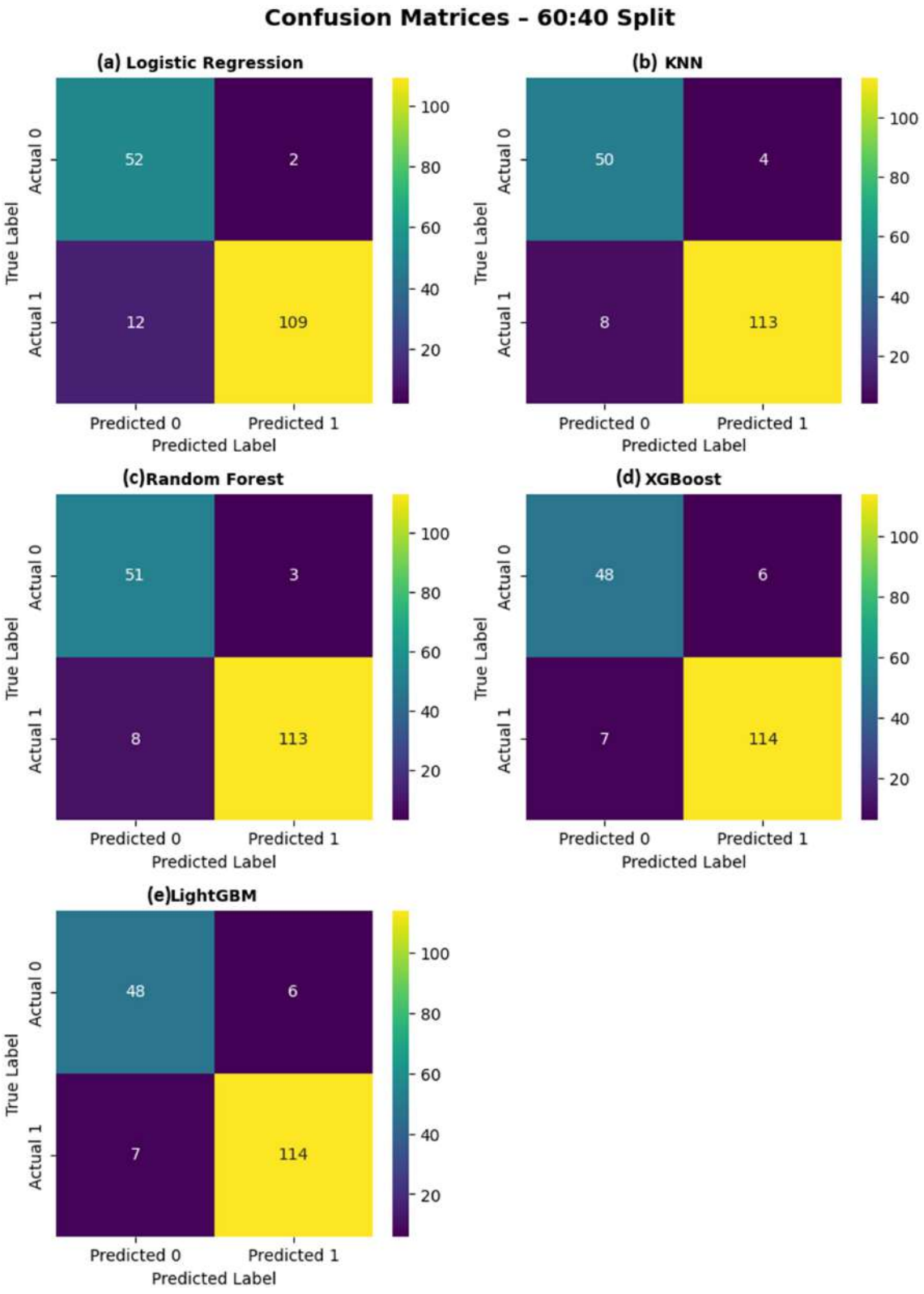

Fig. 6. Confusion matrices of best runs – Stratified 60:40 split.

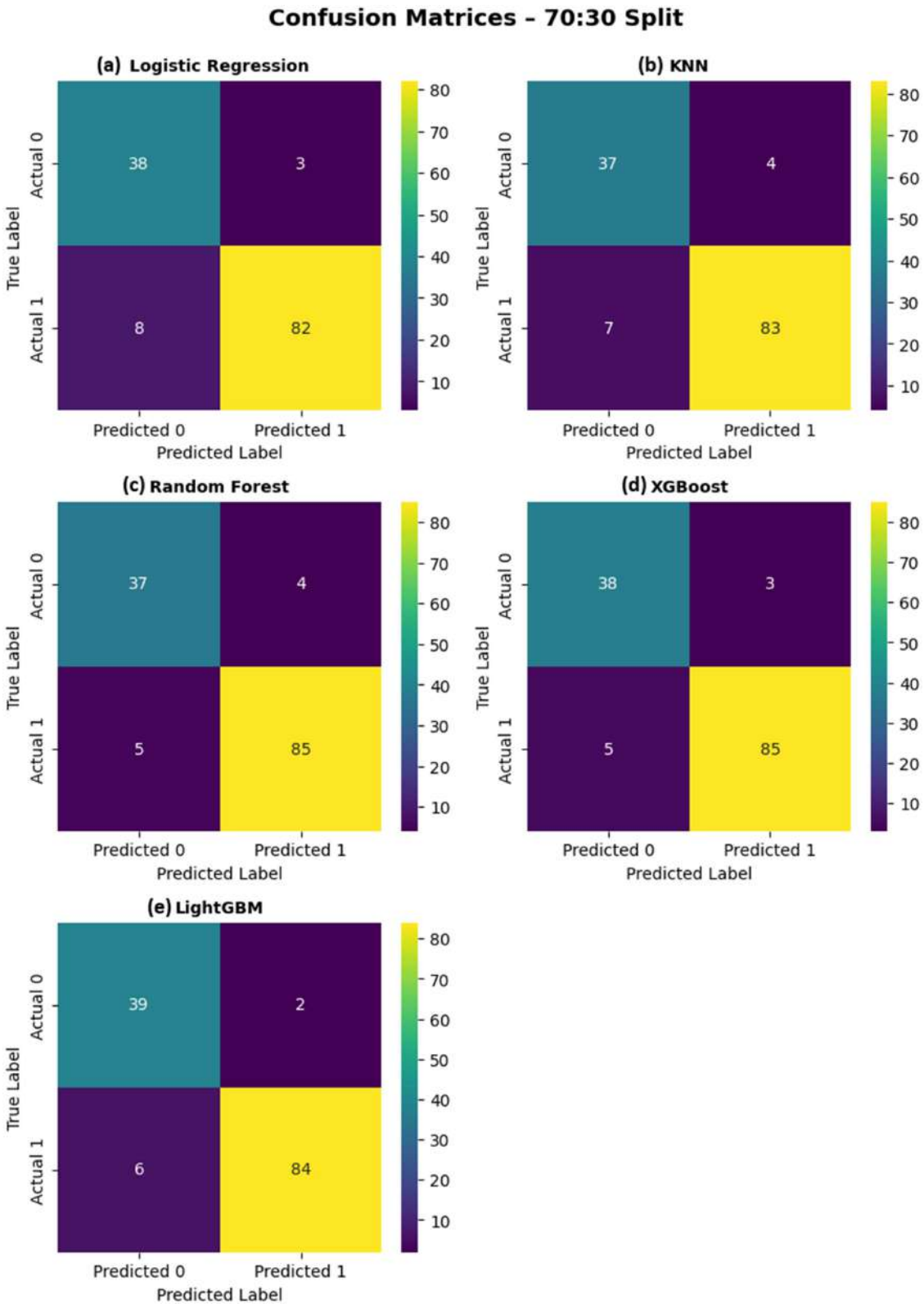

Fig. 7. Confusion matrices of best runs – Stratified 70:30 split.

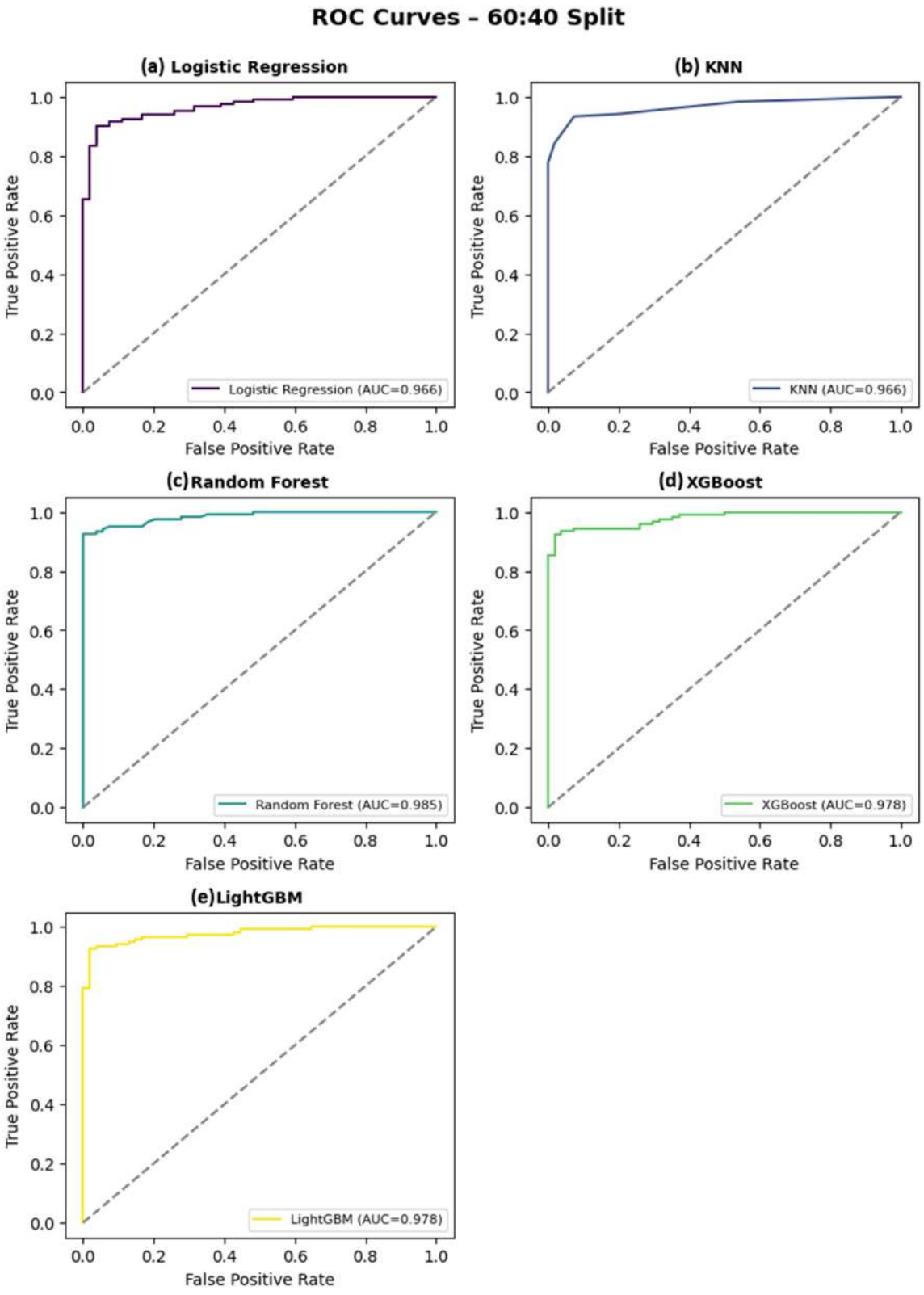

Fig. 8. ROC curves of best runs – Stratified 60:40 split.

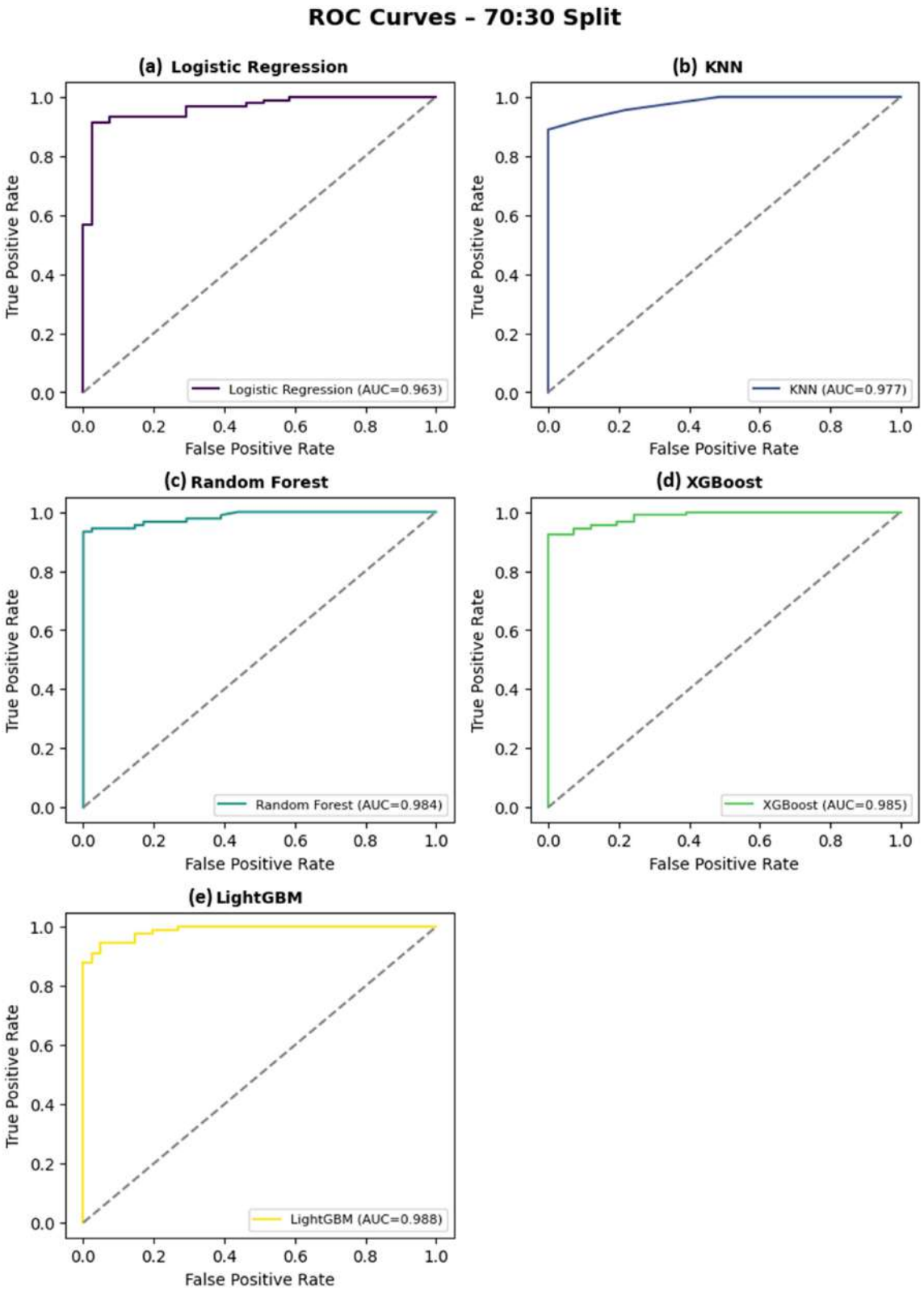

Fig. 9. ROC curves of best runs – Stratified 70:30 split.

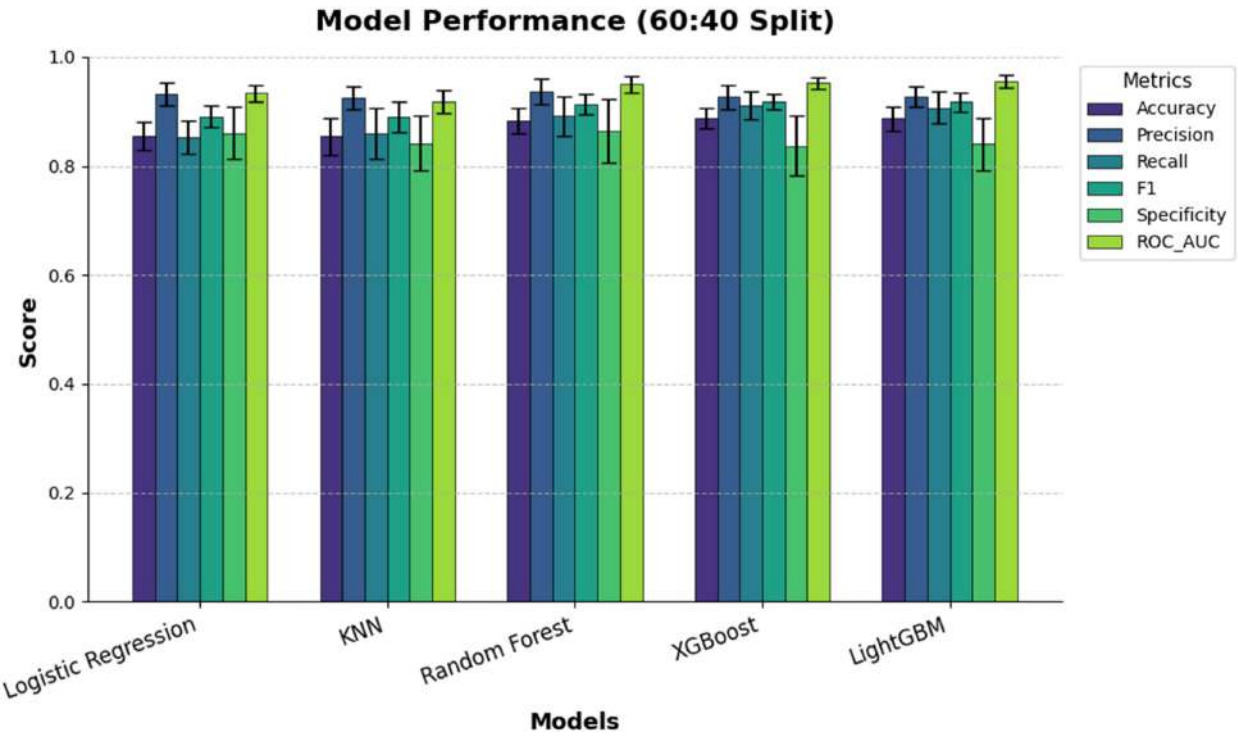

Fig. 10. Comparative model performance across metrics (60:40 Split).

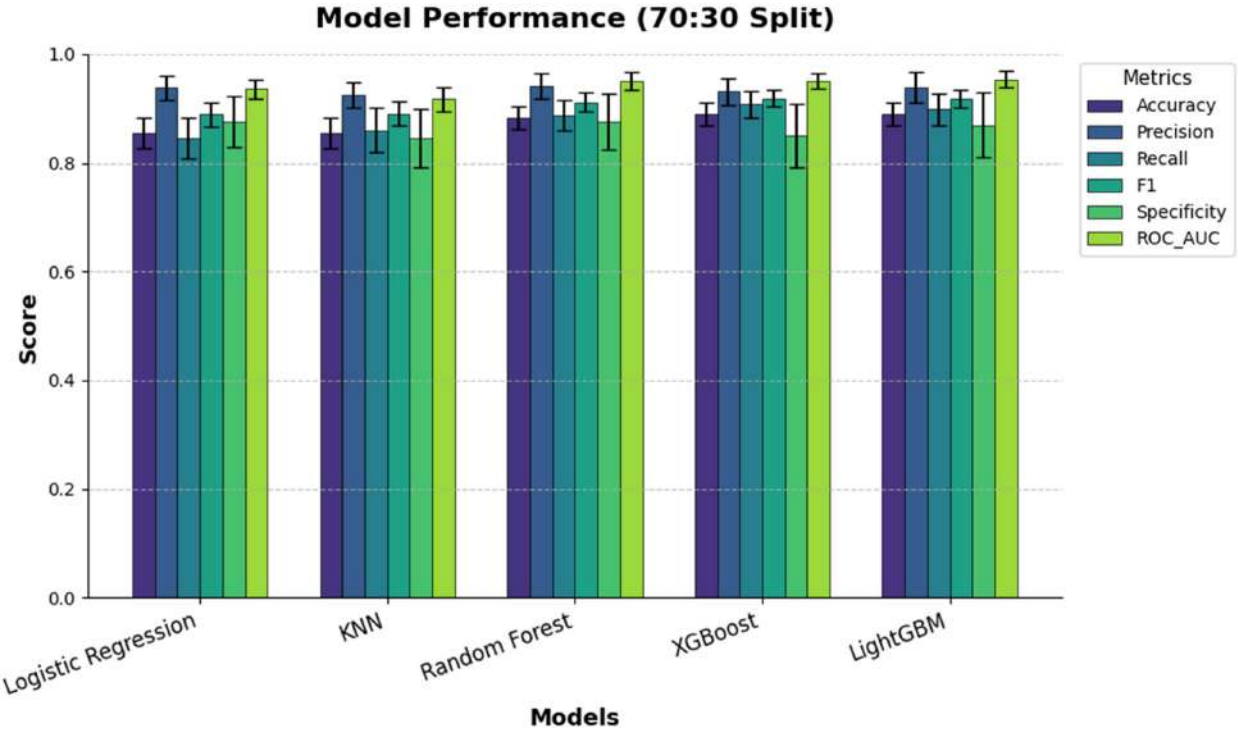

Fig. 11. Comparative model performance across metrics (70:30 Split).

was conducted. Accordingly, interpretability findings are not reported here and will be examined in future extensions.

5. Conclusion

This study demonstrates the effectiveness of advanced machine learning techniques for predicting dementia status using the OASIS cross-sectional MRI dataset. Following rigorous data curation and preprocessing—comprising feature-level missing-value treatment for predictors (with strict preservation of target-label integrity), outlier mitigation, skewness adjustment, and selective input standardization—a robust analytic pipeline was established to support reliable comparative model evaluation. Across repeated stratified experiments, ensemble-based methods consistently achieved the strongest predictive performance. LightGBM provided the most balanced discrimination, achieving an accuracy of 0.890 ± 0.022 and an ROC-AUC of 0.954 ± 0.015 under the 70:30 split, while XGBoost achieved marginally stronger sensitivity with an accuracy of 0.890 ± 0.020 and an ROC-AUC of 0.951 ± 0.014. Random Forest further reinforced the utility of tree-based learners by yielding the highest specificity, thereby reducing false-positive classifications. Collectively, these findings support the value of ensemble learning as a comparative methodological baseline for dementia prediction and highlight performance stability under repeated internal evaluation.

Because evaluation was limited to internal validation using repeated stratified splits from a single dataset, the results should be interpreted as comparative evidence of model performance within the studied cohort rather than confirmed population-level generalizability. Future work will prioritize external validation on independent clinical cohorts to assess real-world transportability and will explore multimodal and longitudinal extensions to better reflect clinical complexity. Future work will include model explainability and clinical interpretability analyses (e.g., feature importance and SHAP-based attribution) to identify the most influential predictors and support transparent decision-making.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. E. Nichols, J. D. Steinmetz, S. E. Vollset, *et al.*, “Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019,” *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022, https://doi.org/10.1016/S2468-2667(21)00249-8.
2. M. Kivimäki, G. Livingston, A. Singh-Manoux, N. Mars, J. V. Lindbohm, J. Pentti, S. T. Nyberg, M. Pirinen, E. L. Anderson, A. D. Hingorani, and P. N. Sipilä, “Estimating dementia risk using multifactorial prediction models,” *JAMA Network Open*, vol. 6, no. 6, p. e2318132, 2023, https://doi.org/10.1001/jamanetworkopen.2023.18132.
3. W. Zhu, H. Tang, H. Zhang, H. R. Rajamohan, S. - L. Huang, X. Ma, … N. Razavian, “Predicting risk of Alzheimer’s diseases and related dementias with an AI foundation model on electronic health records (TRADE),” *medRxiv*, 2024, https://doi.org/10.1101/2024.04.26.24306180.
4. A. S. Tsang, K. Ibrahim, V. Lostanlen, *et al.*, “Leveraging electronic health records and knowledge networks to improve disease risk prediction models,” *Nature Aging*, vol. 4, no. 8, pp. 1056–1070, 2024, https://doi.org/10.1038/s43587-024-00573-8.
5. J. Venugopalan, L. Tong, H. R. Hassanzadeh, *et al.*, “Multimodal deep learning models for early detection of Alzheimer’s disease stage,” *Scientific Reports*, vol. 11, p. 3254, 2021, https://doi.org/10.1038/s41598-020-74399-w.
6. J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “DeepSurv: A Cox proportional hazards deep neural network,” *arXiv*, 2018, https://arxiv.org/abs/1606.00931.
7. S. Yuan, Q. Liu, X. Huang, *et al.*, “Development of an individualized dementia risk prediction model using deep learning survival analysis incorporating genetic and environmental factors,” *Alzheimer’s Research & Therapy*, vol. 16, p. 278, 2024, https://doi.org/10.1186/s13195-024-01663-w.
8. L. A. McGuinness, C. Warren-Gash, L. R. Moorhouse, S. L. Thomas, and S. M. Langan, “The validity of dementia diagnoses in routinely collected electronic health records in the United Kingdom: A systematic review,” *Pharmacoepidemiology and Drug Safety*, vol. 28, no. 2, pp. 244–255, 2019, https://doi.org/10.1002/pds.4669.
9. T. Wilkinson, C. Schnier, K. Bush, K. Rannikmäe, D. E. Henshall, C. Lerpiniere, N. E. Allen, R. Flaig, T. C. Russ, D. Bathgate, S. Pal, J. T. O’Brien, C. L. M. Sudlow, and Dementias Platform UK and UK Biobank, “Identifying dementia outcomes in UK Biobank: A validation study of primary care, hospital admissions and mortality data,” *European Journal of Epidemiology*, vol. 34, no. 6, pp. 557–565, 2019, https://doi.org/10.1007/s10654-019-00499-1.
10. S. E. Davis, R. A. Greevy, Jr., T. A. Lasko, C. G. Walsh, and M. E. Matheny, “Detection of calibration drift in clinical prediction models to inform model updating,” *Journal of Biomedical Informatics*, vol. 112, p. 103611, 2020, https://doi.org/10.1016/j.jbi.2020.103611.
11. A. Subbaswamy and S. Saria, “From development to deployment: Dataset shift, causality, and shift-stable models in health AI,” *Biostatistics*, vol. 21, no. 2, pp. 345–352, 2019, https://doi.org/10.1093/biostatistics/kxz041.
12. S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 10, p. 100804, 2023, https://doi.org/10.1016/j.patter.2023.100804.
13. A. J. Vickers and E. B. Elkin, “Decision curve analysis: A novel method for evaluating prediction models, diagnostic tests,

and molecular markers," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006, https://doi.org/10.1177/0272989$\times$06295361.
14. J. F. Cohen and P. M. Bossuyt, "TRIPOD + AI: An updated reporting guideline for clinical prediction models," *BMJ*, vol. 385, p. q824, 2024, https://doi.org/10.1136/bmj.q824.
15. R. D. Riley, L. Archer, K. I. Snell, J. Ensor, P. Dhiman, G. P. Martin, ... G. S. Collins, "Evaluation of clinical prediction models (part 2): How to undertake an external validation study," *BMJ*, vol. 384, pp. 2023–074820, 2024, https://doi.org/10.1136/bmj-2023-074820.
16. S. M. Lundberg and S. - I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
17. S. Gill, P. Mouches, S. Hu, D. Rajashekar, F. P. MacMaster, E. E. Smith, ... and Alzheimer's Disease Neuroimaging Initiative, "Using machine learning to predict dementia from neuropsychiatric symptom and neuroimaging data," *Journal of Alzheimer's Disease*, vol. 75, no. 1, pp. 277–288, 2020.
18. D. Stamate, R. Smith, R. Tsygancov, R. Vorobev, J. Langham, D. Stahl, and D. Reeves, "Applying deep learning to predicting dementia and mild cognitive impairment," In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, Cham: Springer International Publishing, May 2020, pp. 308–319.
19. J. Kim and J. Lim, "A deep neural network-based method for prediction of dementia using big data," *International Journal of Environmental Research and Public Health*, vol. 18, no. 10, p. 5386, 2021.
20. H. Guram and A. Sharma, "Comparing the performance of various machine learning approaches in identifying different classes of dementia," In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*, April 2021.
21. K. Sakatani and G. Yener, "Application of machine learning in the diagnosis of dementia," *Frontiers in Neurology*, vol. 13, p. 860607, 2022.
22. J. You, Y. R. Zhang, H. F. Wang, M. Yang, J. F. Feng, J. T. Yu, and W. Cheng, "Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study," *EClinicalMedicine*, vol. 53, 2022.
23. S. Li, P. Dexter, Z. Ben-Miled, and M. Boustani, "Dementia risk prediction using decision-focused content selection from medical notes," *Computers in Biology and Medicine*, vol. 182, p. 109144, 2024.
24. A. Zadgaonkar, R. Keskar, and O. Kakde, "Towards a machine learning model for detection of dementia using lifestyle parameters," *Applied Sciences*, vol. 13, no. 19, p. 10630, 2023.
25. Y. Song, Y. Sun, Q. Weng, and L. Yi, "Using machine learning model for predicting risk of memory decline: A cross-sectional study," *Heliyon*, vol. 10, no. 20, 2024.
26. K. Ports, J. Dai, K. Conniff, M. M. Corrada, S. M. Manson, J. O'Connell, and L. Jiang, "Machine learning to predict dementia for American Indian and Alaska Native peoples: A retrospective cohort study," *The Lancet Regional Health–Americas*, vol. 43, 2025.
27. S. Akter, Z. Liu, E. J. Simoes, and P. Rao, "Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's disease and related dementias," *The Journal of Prevention of Alzheimer's Disease*, p. 100169, 2025.
28. Washington University in St. Louis, *Open Access Series of Imaging Studies (OASIS)*, 2025. Retrieved from https://sites.wustl.edu/oasisbrains/.
29. M. Rashidi, S. Arima, A. C. Stetco, C. Coppola, D. Musarò, M. Greco, and M. Maffia, "Prediction of Parkinson Disease Using Long-Term, Short-Term Acoustic Features Based on Machine Learning," *Brain Sci.*, vol. 15, p. 739, 2025, https://doi.org/10.3390/brainsci15070739.
30. S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of Machine Learning Methods in Diagnosing Parkinson's Disease Based on Dysphonia Measures," *Biomed. Eng. Lett.*, vol. 8, pp. 29–39, 2018, https://doi.org/10.1007/s13534-017-0058-2.
31. V. Škvára, T. Pevný, and V. Šmídl, "Is AUC the best measure for practical comparison of anomaly detectors?," *arXiv preprint arXiv:2305.04754*, 2023.