

When Fewer Variables Change More Than Accuracy: Model-Specific Generalization and Error Trade-Offs in Pre-Diagnostic Alzheimer's Prediction from Routine Health Records

Emma Wilson^{1*}, Frederik De Jong², Lara Peters²

¹ Department of Neurology and AI Diagnostics, Faculty of Medicine, University of Edinburgh, Edinburgh, United Kingdom.

² Department of Integrative Healthcare Data and Predictive Analytics, Faculty of Medicine, Utrecht University, Utrecht, Netherlands.

*E-mail ✉ emma.wilson@ed.ac.uk

Abstract

High-dimensional routine healthcare records provide an opportunity to identify individuals at elevated risk of Alzheimer's disease before formal diagnosis, but the value of thousands of available variables cannot be judged from predictive discrimination alone. This study re-examines an existing machine-learning experiment to determine how aggressive feature-space reduction affects discrimination, class-balanced performance, and model behavior when Alzheimer's disease is predicted 3–4 years before clinical diagnosis. Integrated primary-healthcare and social-service records were used to construct a 4,311-feature modelling space. Analysis of variance, mutual information, and SHapley Additive exPlanations were used to generate 25- and 50-feature subsets, which were evaluated against the complete feature space across Logistic Regression, Random Forest, Gaussian Naïve Bayes, Multi-Layer Perceptron, XGBoost, and LightGBM. Feature compression produced strongly model-dependent effects. The highest AUC was 0.755 for XGBoost using 50 SHAP-selected variables, whereas the highest balanced accuracy was 0.668 for Logistic Regression using the same feature count and selection strategy. Several compact configurations approached or exceeded the discrimination achieved with all 4,311 variables, but substantial deterioration occurred in other model-selection combinations, particularly Gaussian Naïve Bayes with SHAP-selected features. These findings indicate that dimensionality reduction cannot be evaluated as a single universally beneficial preprocessing step. Compact Alzheimer's prediction models should instead be judged according to how much discrimination they retain, whether balanced performance improves or deteriorates, how strongly results depend on the learning algorithm, and whether reduced feature spaces support more stable and interpretable modelling.

Keywords: Alzheimer's disease, Electronic health records, Feature compression, Machine learning, SHAP, Model generalization

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder whose clinical consequences

extend beyond cognitive decline to substantial patient, caregiver, healthcare, and social-service burden [1]. The possibility of identifying individuals at elevated risk before formal diagnosis has therefore motivated increasing interest in prediction from routinely collected healthcare information. Real-world electronic health records (EHRs) are particularly attractive because they are generated continuously during ordinary care and may contain diagnostic, medication, laboratory, utilization, and functional signals long before specialist assessment occurs [2]. Early prediction has also been linked to opportunities for earlier supportive intervention, risk

Access this article online

<https://smerpub.com/>

Received: 25 September 2023; Accepted: 14 December 2023

Copyright CC BY-NC-SA 4.0

How to cite this article: Wilson E, De Jong F, Peters L. When Fewer Variables Change More Than Accuracy: Model-Specific Generalization and Error Trade-Offs in Pre-Diagnostic Alzheimer's Prediction from Routine Health Records. J Med Sci Interdiscip Res. 2023;3(2):184-98. <https://doi.org/10.51847/HrOcaDzUxJ>

modification, and clinical assessment, although prediction itself must not be equated with diagnosis or proven disease modification [3-5].

Machine-learning approaches have increasingly been applied to dementia and AD prediction using multimodal data, administrative health records, claims, imaging, and longitudinal clinical information. Deep-learning studies have demonstrated that complex representations derived from imaging and EHR sources can contribute to disease-stage classification and progression modelling [6, 7]. Register-based work has likewise shown that routinely accumulated healthcare information may contain longitudinal patterns relevant to later AD diagnosis [8]. These developments make routine clinical data an appealing alternative or complement to specialized biomarkers, particularly where access to advanced diagnostic procedures is limited.

The principal difficulty is dimensionality. Modern EHR systems may encode thousands of potentially relevant variables across diagnoses, medications, laboratory orders, observations, healthcare contacts, procedures, home-care records, and other administrative traces [9]. A larger feature space can preserve more potentially useful information, but it also introduces redundancy, sparsity, correlated variables, computational burden, and increased difficulty in interpreting fitted models. Feature selection is commonly proposed as a way to reduce these problems by restricting prediction to variables that carry the greatest information for the target outcome [10]. The optimal reduction strategy, however, cannot be assumed to be independent of the prediction algorithm because different learning methods exploit linear associations, nonlinearities, interactions, thresholds, and correlated inputs differently [11].

Existing AD-prediction studies illustrate this methodological diversity. EHR-based cohort-discovery and risk-stratification work has used machine-learning approaches to identify individuals at elevated AD risk [12]. Sparse logistic modelling has shown that relatively constrained sets of claims-derived variables can support dementia prediction [13], while ensemble approaches using cognitive features have also applied explicit feature reduction before classification [14]. Other studies have emphasized explainability and deep architectures for understanding disease progression [15], and filter- and wrapper-based approaches have been used within models predicting AD or mild cognitive impairment trajectories [16]. Collectively, these studies establish that variable selection can be important, but they do not imply that

reducing the feature space produces the same effect across model families or performance metrics.

A further issue is that feature-selection studies are often interpreted primarily through headline discrimination. A reduced model may be described as superior because its AUC increases, yet a similar AUC can coexist with materially different balanced accuracy, sensitivity, specificity, generalization behavior, or distribution of classification errors. Conversely, a feature set may retain nearly all of the discrimination of a much larger model while offering substantially lower dimensionality even if its AUC does not numerically exceed the full-feature baseline. Evaluating feature compression therefore requires examining what information is retained, what information is lost, and how different algorithms react to the reduced representation.

The present study reframes an existing AD-prediction experiment around this question. Rather than treating ANOVA, mutual information, and SHAP primarily as competing feature-ranking techniques, the analysis examines their 25- and 50-variable subsets as controlled perturbations of a 4,311-variable modelling space. The objective is to determine how aggressive feature-space reduction affects discrimination, class-balanced performance, generalization, and classification-error profiles across six machine-learning algorithms predicting AD from longitudinal routine healthcare and social-service records 3–4 years before diagnosis.

Materials and Methods

Study population and integrated health/social-care data

The source dataset comprised integrated records from the Primary Healthcare and Social Service systems of Kuopio, Finland, covering service use between 2012 and 2020. The initial population included 26,828 residents who were aged at least 65 years in 2015. In this broader source population, 58% were female and 42% were male, and the mean age was 76 years.

The information architecture was substantially broader than a conventional disease registry. Primary-healthcare information was distributed across 25 tables containing demographic information, treatment periods, ICD-10 diagnoses, International Classification of Primary Care (ICPC) codes, visit characteristics, home-care interruptions, medication records, laboratory requests and results, memory tests, service and care plans, risk information, vaccination records, imaging referrals, procedure codes, and physiological observations. Six

additional social-service tables described process initiation, assessments, decisions, allocated services, payments, and payment liabilities. Secondary processing of these health and social data was conducted in a controlled audited environment consistent with Finnish requirements governing secondary use of health and social information [17].

The distinction between the broader source population and the final modelling cohort was maintained throughout analysis. The larger 26,828-person population characterized the available healthcare data resource, whereas subsequent eligibility criteria and AD/control construction produced the cohort used for machine-learning evaluation.

Table 1. Dataset and experimental design used for feature-compression analysis

Component	Source-locked specification
Source population	26,828 adults aged ≥ 65 years
Geographic setting	Kuopio, Finland
Source-data period	2012–2020
Primary-healthcare structure	25 source tables
Social-service structure	6 source tables
Final modelling controls	8,005
Final modelling AD cases	1,524
Final modelling cohort ratio	Approximately 5 controls:1 AD case
Final modelling cohort mean age	78 years
Final modelling cohort female proportion	66%
Candidate feature space	4,311 variables
Prediction horizon	Approximately 3–4 years before diagnosis
Feature-selection methods	ANOVA, mutual information, SHAP
Reduced feature spaces	Top 25 and top 50
Full-feature comparator	All 4,311 variables
Machine-learning algorithms	LR, RF, GNB, MLP, XGBoost, LGBM
Train/test partition	75% / 25%
Hyperparameter optimization	Grid Search
Cross-validation	Stratified 10-fold CV on training data
Model-selection metric	AUC
Final evaluation metrics	AUC and balanced accuracy with 95% CIs

Cohort definition and temporal prediction window

AD status was defined using ICD-10 codes G30 and F00. To ensure that models were built from sufficiently long pre-diagnostic records, individuals with AD were required to have at least four years of available historical data before the recorded diagnosis date. For controls, individuals with documented memory disorders were excluded, followed by additional removal of those with dementia-related medication records or diagnostic codes indicating mild cognitive impairment, unspecified dementia, dementia, or memory problems.

Eligible controls were subsequently matched to the AD cohort on age and sex. Primary-healthcare and social-service information was then integrated using patient identifiers. The final modelling population comprised

8,005 controls and 1,524 AD cases, corresponding to an approximately 5:1 ratio. The final cohort was 66% female with a mean age of 78 years.

The longitudinal design was constructed specifically to avoid treating information recorded immediately before diagnosis as genuinely early predictive evidence. Features were generated over four years of historical records using six-month intervals, but the six months immediately preceding the diagnosis date were excluded. The modelling task therefore concentrated on information arising approximately 3–4 years before formal AD diagnosis.

Control participants were assigned pseudo-diagnosis dates that ensured an equivalent antecedent observation period. The resulting design allowed the same temporal

feature-construction framework to be applied to AD and control participants while reducing the risk that overt near-diagnostic manifestations dominated the prediction problem.

High-dimensional feature construction

Data preprocessing consisted of cleaning, temporal feature engineering, categorical encoding, and missing-data control. Duplicate or erroneous records were removed before construction of the modelling variables. Categorical events were converted through one-hot encoding, whereas numerical features summarized the frequency or magnitude of recorded events within the prespecified temporal windows.

The resulting variables represented multiple dimensions of healthcare use and clinical status, including diagnostic codes, medication categories, numbers and types of clinical contacts, professional resources involved in visits, laboratory requests, observations, memory testing, service and care plans, and selected social-service activity. Columns with substantial missingness were removed instead of applying imputation.

After preprocessing, the modelling matrix contained 4,311 candidate features. This full representation constituted the baseline against which aggressive feature compression was evaluated.

Feature compression strategies

Three feature-selection approaches were used to construct reduced representations of the 4,311-variable matrix.

Analysis of variance (ANOVA) ranked individual variables according to their capacity to differentiate the two outcome classes through the F statistic. The method evaluates whether between-class variation is large relative to within-class variation and therefore provides a univariate filter before classifier fitting [18]. Its use as a feature-screening strategy has been described in machine-learning applications involving high-dimensional biomedical data [19].

Mutual information (MI) was used as a second pre-classifier filter. MI measures the shared information between a predictor and the outcome and can capture statistical dependence without restricting the relationship to simple linear association [20]. Feature selection based on MI can therefore identify variables that contribute information about the target while also providing a model-independent ranking before classifier training [21].

SHapley Additive exPlanations (SHAP) represented the post-classifier, model-dependent strategy. SHAP attributes a model prediction to individual variables using principles derived from Shapley values [22]. The underlying allocation concept originates from cooperative game theory [23]. For the tree-based models in the source experiment, features were ranked using their mean absolute SHAP values across the dataset, enabling global importance rankings derived from the fitted model itself [24].

For each selection strategy, two reduced feature spaces were constructed: the top 25 variables and the top 50 variables. These were evaluated against the complete 4,311-feature space. Thus, feature compression reduced the available representation to approximately 0.58% or 1.16% of the original variable count, respectively. These percentages are simple arithmetic descriptions of the source-reported feature counts rather than independently analysed outcomes.

Machine-learning models

Six classifiers were evaluated to determine whether the effect of compression depended on model family.

Logistic Regression (LR) provided a linear probabilistic baseline in which input variables are combined through a logistic transformation to estimate class probability [25]. Its use in health-prediction settings provides a comparatively interpretable reference against more complex nonlinear models [26]. LR has also been combined with feature-selection procedures in AD classification research [27].

Gaussian Naïve Bayes (GNB) applies Bayes' theorem under a conditional-independence assumption and models continuous predictors using Gaussian distributions [28]. Its computational simplicity has supported its application across multi-disease classification problems [29], including health datasets in which efficient probabilistic classification is desirable [30].

Random Forest (RF) is an ensemble method that aggregates predictions from multiple decision trees constructed using bootstrap samples and randomly selected predictor subsets [31]. The resulting diversity among trees can improve classification robustness relative to individual decision trees [32].

The Multi-Layer Perceptron (MLP) represented a feedforward neural-network classifier capable of learning nonlinear combinations of input variables [33]. Such multilayer architectures have been applied to

health-risk classification where interactions among variables may not conform to simple linear structure [34]. XGBoost was included as a gradient-boosted tree ensemble in which sequential trees are optimized to correct errors from preceding learners [35]. LightGBM (LGBM) provided a second gradient-boosting implementation designed for computational efficiency on large feature spaces through techniques such as Gradient-based One-Side Sampling and Exclusive Feature Bundling [36].

The diversity of these six algorithms was central to the compression analysis because a feature subset retained by one selection strategy was not assumed to have equivalent utility for all model classes.

Training, validation, and performance evaluation

The complete dataset was divided into a 75% training set and a 25% independent test set. Continuous variables were standardized before model fitting. Hyperparameters were selected through Grid Search combined with stratified 10-fold cross-validation performed only on the training data.

AUC was used as the optimization criterion during model and hyperparameter selection. Final performance on the independent test set was evaluated using both AUC and balanced accuracy. Balanced accuracy was particularly relevant because the final cohort was imbalanced, and the metric gives equal weight to sensitivity and specificity rather than allowing the majority class to dominate the summary measure [37].

The experimental design therefore produced, for each classifier, results for the complete 4,311-feature representation and for multiple 25- and 50-feature configurations generated through ANOVA, MI, and SHAP.

Results and Discussion

Predictive performance across full and reduced feature spaces

Feature compression did not exert a uniform effect across the six algorithms. The highest AUC observed across all tested configurations was 0.755 [0.725, 0.785], obtained by XGBoost using the top 50 SHAP-selected features. The highest balanced accuracy was instead produced by

Logistic Regression with the same number and selection type of variables, reaching 0.668 with a source-reported confidence interval rendered as [0.642, 694]. Because the upper bound is malformed in the source table, the point estimate is retained as reported and the confidence-interval issue is treated as a transcription inconsistency rather than silently presented as independently verified.

The complete 4,311-feature models remained competitive for several algorithms. LR achieved its highest AUC, 0.751 [0.726, 0.776], with the complete feature set rather than a reduced SHAP model. XGBoost likewise produced an AUC of 0.753 [0.728, 0.778] with all variables, only 0.002 below its maximum with 50 SHAP-selected features. For MLP, full-dimensional modelling produced a balanced accuracy of 0.615 [0.595, 0.636], which exceeded every reduced-feature MLP configuration. These findings show that dimensionality reduction cannot be characterized as uniformly performance-enhancing.

At the same time, substantial compression often preserved discrimination remarkably well. LR using only 25 ANOVA-ranked variables achieved an AUC of 0.748 [0.721, 0.775], close to its complete-feature AUC of 0.751. LGBM reached 0.754 [0.729, 0.779] with only 25 SHAP-selected variables, compared with 0.743 [0.718, 0.768] using all 4,311 variables. RF obtained its highest AUC of 0.732 [0.707, 0.757] with 25 SHAP-selected variables, while its complete-feature AUC was 0.728 [0.703, 0.753].

The opposite pattern was evident for GNB under SHAP compression. Its complete-feature AUC was 0.621 [0.596, 0.646], whereas SHAP reduction decreased AUC to 0.567 [0.544, 0.590] with 25 variables and 0.561 [0.538, 0.584] with 50. Balanced accuracy simultaneously fell from 0.602 [0.582, 0.623] using all variables to 0.497 under both SHAP-reduced configurations. In contrast, GNB reached its best AUC of 0.723 [0.698, 0.748] with the 25 ANOVA-selected variables. The direction of the effect therefore depended not only on the number of retained variables but on the interaction between the selection method and the classifier.

Table 2 presents the complete source-locked performance matrix.

Table 2. AUC and balanced accuracy across full and compressed feature spaces

Feature condition	Metric	LR	RF	GNB	MLP	XGBoost	LGBM
ANOVA — top 25	Balanced accuracy	0.603 [0.584, 0.622]	0.608 [0.590, 0.629]	0.612 [0.591, 0.633]	0.594 [0.574, 0.615]	0.612 [0.592, 0.631]	0.600 [0.581, 0.621]
ANOVA — top 25	AUC	0.748 [0.721, 0.775]	0.710 [0.685, 0.735]	0.723 [0.698, 0.748]	0.738 [0.713, 0.763]	0.743 [0.718, 0.768]	0.730 [0.705, 0.755]
ANOVA — top 50	Balanced accuracy	0.601 [0.581, 0.621]	0.604 [0.585, 0.624]	0.612 [0.589, 0.635]	0.600 [0.581, 0.620]	0.614 [0.595, 0.634]	0.609 [0.590, 0.629]
ANOVA — top 50	AUC	0.741 [0.714, 0.768]	0.720 [0.695, 0.745]	0.685 [0.654, 0.716]	0.700 [0.675, 0.725]	0.744 [0.719, 0.769]	0.732 [0.707, 0.757]
MI — top 25	Balanced accuracy	0.592 [0.573, 0.612]	0.598 [0.579, 0.618]	0.599 [0.580, 0.619]	0.589 [0.569, 0.610]	0.604 [0.569, 0.610]	0.607 [0.588, 0.627]
MI — top 25	AUC	0.736 [0.709, 0.763]	0.708 [0.683, 0.733]	0.709 [0.684, 0.734]	0.716 [0.691, 0.741]	0.744 [0.719, 0.769]	0.735 [0.710, 0.760]
MI — top 50	Balanced accuracy	0.600 [0.581, 0.621]	0.603 [0.584, 0.624]	0.604 [0.585, 0.623]	0.596 [0.576, 0.617]	0.612 [0.593, 0.632]	0.609 [0.590, 0.630]
MI — top 50	AUC	0.739 [0.712, 0.766]	0.724 [0.699, 0.749]	0.702 [0.677, 0.727]	0.699 [0.674, 0.724]	0.749 [0.724, 0.774]	0.740 [0.715, 0.765]
SHAP — top 25	Balanced accuracy	0.662 [0.640, 0.684]	0.589 [0.567, 0.610]	0.497 [0.476, 0.520]	0.538 [0.517, 0.560]	0.597 [0.576, 0.618]	0.607 [0.588, 0.627]
SHAP — top 25	AUC	0.723 [0.698, 0.748]	0.732 [0.707, 0.757]	0.567 [0.544, 0.590]	0.730 [0.705, 0.755]	0.748 [0.723, 0.773]	0.754 [0.729, 0.779]
SHAP — top 50	Balanced accuracy	0.668 [0.642, 694]*	0.580 [0.555, 0.605]	0.497 [0.472, 0.521]	0.545 [0.520, 0.570]	0.616 [0.594, 0.638]	0.592 [0.571, 0.614]
SHAP — top 50	AUC	0.727 [0.697, 0.757]	0.724 [0.699, 0.749]	0.561 [0.538, 0.584]	0.726 [0.701, 0.751]	0.755 [0.725, 0.785]	0.751 [0.726, 0.776]
Complete 4,311	Balanced accuracy	0.647 [0.625, 0.668]	0.591 [0.571, 0.611]	0.602 [0.582, 0.623]	0.615 [0.595, 0.636]	0.627 [0.607, 0.647]	0.615 [0.595, 0.635]
Complete 4,311	AUC	0.751 [0.726, 0.776]	0.728 [0.703, 0.753]	0.621 [0.596, 0.646]	0.650 [0.625, 0.675]	0.753 [0.728, 0.778]	0.743 [0.718, 0.768]

*The LR SHAP-50 balanced-accuracy CI is reproduced as printed in the source and contains a malformed upper bound. No external correction was imposed.

How strongly can 4,311 features be compressed without losing discrimination?

The observed results show that an extremely small proportion of the original feature space could retain most, and in selected configurations all, of the discriminative performance of the full-dimensional models. Fifty variables represent approximately 1.16% of the 4,311-variable space, while 25 variables represent

approximately 0.58%. Despite this reduction, multiple compact configurations produced AUCs within a few thousandths of, or above, their corresponding complete-feature models.

This compression effect was particularly clear for the boosting models. XGBoost increased from an AUC of 0.753 with all features to 0.755 with 50 SHAP-ranked variables. LGBM increased from 0.743 with the

complete representation to 0.754 with only 25 SHAP-selected features. RF similarly rose from 0.728 with all variables to 0.732 with SHAP-25. These differences were numerically modest, but the dimensionality reduction was substantial.

LR demonstrated a different pattern. Its maximum discrimination remained associated with the full 4,311-variable representation, AUC 0.751. Nevertheless, reducing the feature space to 25 ANOVA-ranked variables produced an AUC of 0.748, indicating that nearly all of the observed discrimination could be retained with fewer than 1% of the candidate variables. The same compression did not preserve every performance dimension equally: LR's balanced accuracy increased markedly under SHAP selection even though its AUC decreased relative to the complete-feature model.

MLP also illustrates why compression cannot be interpreted from AUC alone. Its complete-feature AUC was 0.650, whereas ANOVA-25 yielded 0.738 and SHAP-25 yielded 0.730. Yet its balanced accuracy was highest with all 4,311 variables at 0.615, compared with 0.594 under ANOVA-25 and 0.538 under SHAP-25. Thus, feature reduction could improve ranking discrimination while simultaneously reducing class-balanced classification performance.

The contrast across these models motivates a performance-retention view of dimensionality reduction rather than a binary "selection helps/selection harms" judgment.

Model-specific responses to ANOVA, mutual information, and SHAP selection

ANOVA, MI, and SHAP produced distinct patterns across classifier families. ANOVA was particularly effective for GNB and MLP when only 25 variables were retained. GNB achieved its highest AUC, 0.723, under ANOVA-25, substantially exceeding its complete-feature AUC of 0.621. MLP likewise reached its highest AUC of 0.738 with ANOVA-25, compared with 0.650 using all features. Increasing the ANOVA subset from 25 to 50 did not improve either model's discrimination; GNB declined to 0.685 and MLP to 0.700.

For LR, ANOVA also preserved high discrimination. The top-25 subset produced AUC 0.748, only 0.003 below the complete model. Expanding to 50 ANOVA features reduced AUC slightly further to 0.741. Balanced accuracy remained around 0.60 under both configurations, below the complete-feature value of

0.647. Consequently, ANOVA compression retained LR's ranking performance more effectively than its class-balanced performance.

MI generated comparatively moderate effects. For XGBoost, AUC progressed from 0.744 with MI-25 to 0.749 with MI-50, approaching the full-feature AUC of 0.753. LGBM similarly increased from 0.735 to 0.740 as the MI subset expanded from 25 to 50 variables. RF reached AUC 0.724 with MI-50 compared with 0.728 using all features. These patterns indicate that MI could produce compact models with relatively limited loss of discrimination, although it rarely generated the strongest configuration for a given classifier.

SHAP produced the largest cross-model divergence. Its strongest results occurred for XGBoost, LGBM, RF, and LR depending on the metric examined. XGBoost achieved the study-wide maximum AUC of 0.755 with SHAP-50, while LGBM achieved 0.754 with SHAP-25. RF reached its own maximum AUC of 0.732 with SHAP-25. LR, however, displayed a metric trade-off: SHAP-50 produced the highest balanced accuracy in the entire experiment, 0.668, while its AUC fell to 0.727 from the complete-feature value of 0.751.

The same SHAP reduction that benefited selected models was detrimental to GNB. Both its discrimination and balanced accuracy deteriorated markedly, with AUC values of 0.567 and 0.561 for 25 and 50 SHAP features, respectively, and balanced accuracy of 0.497 in both conditions. MLP also lost balanced accuracy under SHAP despite gaining AUC relative to the complete-feature model.

These results reject the interpretation of feature importance as universally transferable across learning algorithms. The empirically strongest feature set depended on both the classifier and the performance criterion. A subset that improved a boosting model could be poorly suited to a probabilistic classifier, while a configuration that maximized AUC did not necessarily maximize balanced accuracy.

Generalization of the highest-performing compact models

The performance matrix identified XGBoost and Logistic Regression using 50 SHAP-selected variables as particularly informative configurations, but for different reasons. XGBoost produced the highest AUC in the experiment, whereas LR produced the highest balanced accuracy. Learning-curve analysis showed that these

headline results were accompanied by materially different generalization behavior.

For XGBoost with 50 SHAP-selected features, training performance was extremely high when only 10% or 20% of the available training observations were used. Performance on unseen data was considerably weaker at these small training fractions. As progressively larger portions of the training set were introduced, training performance declined modestly while test performance improved. The convergence of the two curves indicates that additional data reduced overfitting. Nevertheless, a persistent separation remained between training and test performance, indicating that the highest-AUC compact model continued to exhibit evidence of overfitting.

LR with the same 50 SHAP-selected variables followed a different trajectory. At the smallest training fractions, training AUC was again substantially higher than test AUC, showing that the linear model was not immune to overfitting when fitted with limited data. As training size increased, however, the difference between training and test performance progressively narrowed. The resulting learning curve was more stable than that observed for XGBoost, with stronger convergence between training and test performance as sample size increased.

These patterns add an important dimension to the comparison of compact models. XGBoost achieved the strongest discrimination on the independent test set, but its learning curve suggested greater sensitivity to training-sample size and residual overfitting. LR did not achieve the highest AUC, yet its learning behavior indicated comparatively more stable generalization with increasing data. Consequently, selection of a compact model cannot be based solely on the single maximum test-set AUC.

Contrasting error profiles of XGBoost and logistic regression

The difference between XGBoost and LR became more consequential when prediction errors were examined directly. The independent test set contained 381 individuals with AD and 2,002 controls. Despite XGBoost having the highest AUC, its classification errors were heavily concentrated among AD cases.

XGBoost with 50 SHAP-selected features correctly identified 92 of the 381 AD cases and misclassified 289 AD cases as controls. Only 17 controls were incorrectly classified as AD. Thus, the model strongly favored correct identification of the control class while failing to identify a large proportion of individuals who subsequently received an AD diagnosis.

LR produced the opposite error pattern. With the same 50-feature SHAP configuration, it correctly classified 247 AD cases and missed 134. This substantial increase in AD detection was accompanied by 625 control participants being incorrectly labelled as AD. LR therefore captured substantially more of the positive class but produced a much larger false-positive burden.

The contrast shows why neither the highest AUC nor the highest balanced accuracy should be interpreted in isolation as evidence of universal superiority. XGBoost generated very few false-positive classifications among controls but missed many AD cases. LR detected many more AD cases while incorrectly classifying substantially more controls. These are qualitatively different model behaviors even though both configurations emerged as leading models under the aggregate evaluation metrics.

Table 3. Classification-error profiles of the leading 50-feature SHAP models

Error-profile component	XGBoost + SHAP-50	Logistic Regression + SHAP-50
Total AD cases in test set	381	381
AD correctly identified	92	247
AD missed and classified as control	289	134
Total controls in test set	2,002	2,002
Controls incorrectly classified as AD	17	625
Controls correctly classified*	1,985	1,377
Dominant classification pattern	Very few false positives but many missed AD cases	More AD cases detected but substantially more false positives
Source-supported interpretation	High specificity, low sensitivity for AD	Higher sensitivity, poorer specificity

*Correct-control counts are arithmetic derivations from the reported total controls and false-positive counts; they were not separately reported as new analyses.

The error profiles are displayed graphically in **Figure 1** to distinguish aggregate performance from the distribution of classification mistakes.

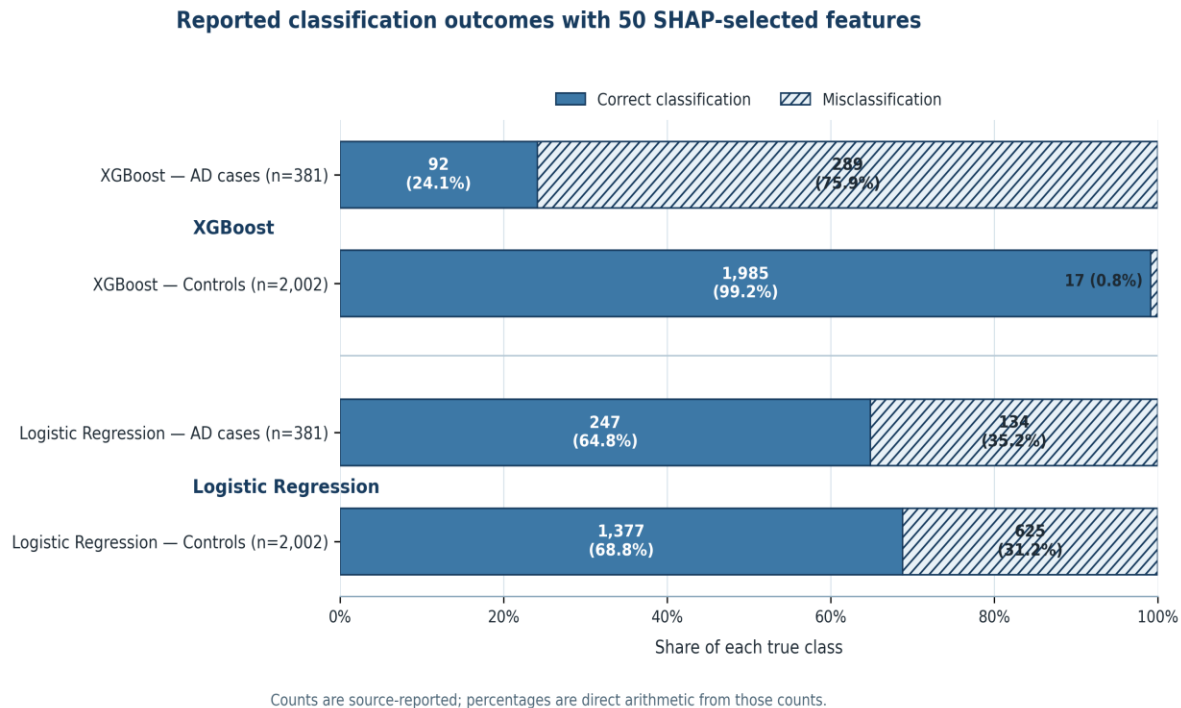


Figure 1. Divergent error profiles of XGBoost and Logistic Regression after reduction to 50 SHAP-selected variables.

Paired error fields show classification outcomes for 381 AD cases and 2,002 controls in the independent test set.

XGBoost correctly identifies 92 AD cases, misses 289, and falsely identifies 17 controls as AD. Logistic

Regression correctly identifies 247 AD cases, misses 134, and falsely identifies 625 controls as AD. Bar proportions are direct arithmetic from the reported confusion-matrix counts. The figure emphasizes that the two leading compact models produce materially different error distributions.

Predictive information retained in the selected features

Feature compression also clarified which categories of routine information remained influential when thousands of candidate variables were reduced to small subsets. Across the pre-classifier rankings, age was consistently one of the strongest discriminators. Diagnostic and symptom-related variables associated with memory impairment, dementia, and mild cognitive impairment also appeared prominently.

ANOVA and MI rankings included ICPC and ICD-10 variables related to dementia and cognitive impairment, home-care activity, service and care planning, memory assessments, follow-up encounters, and medication use. The post-classifier SHAP rankings similarly retained age

and memory-impairment information but also emphasized patterns of healthcare utilization.

For XGBoost, the highest SHAP-ranked feature was age, followed by residential zip code and an ICPC memory-impairment indicator. Other leading variables included the number of medications, numbers and modes of healthcare contacts, laboratory requests, contacts with nurses and physicians, and cardiovascular and infectious-disease medication categories. LGBM yielded a broadly similar pattern, with age, memory impairment, zip code, medication counts, laboratory requests, and healthcare-contact variables appearing prominently.

These rankings should not be interpreted as causal estimates of AD risk. A variable may receive high predictive importance because it captures clinical

manifestations, multimorbidity, healthcare engagement, diagnostic activity, or correlated patterns in the dataset. Feature importance therefore describes contribution to the fitted prediction model, not proof that changing the variable would alter disease risk.

One notable result was the comparatively limited presence of social-service variables among the strongest-ranked predictors. Although the study deliberately integrated primary-healthcare and social-service records, the leading selected features were derived mainly from primary healthcare. Several explanations remain possible. Primary-healthcare records may have contained greater clinical granularity, relevant social-service activity may have occurred later in the disease trajectory, or the available social-service variables may have required representations different from those used in the present modelling pipeline. The finding therefore does not establish that social-service information lacks value; it shows only that such variables were not prominent among the strongest predictors under the tested feature-construction and selection procedures.

The reframed analysis reveals that aggressive reduction of a 4,311-variable routine-healthcare feature space can preserve substantial predictive information, but the consequences of compression depend strongly on the learning algorithm, selection method, and performance criterion. Compact models using only 25 or 50 variables sometimes approached or exceeded their complete-feature counterparts, yet the same form of reduction could also produce substantial deterioration. The central empirical finding is therefore not that fewer variables are inherently superior. It is that dimensionality reduction changes model behavior in ways that cannot be understood from feature count or AUC alone.

Feature compression does not affect all models equally

A central assumption behind feature selection is that removing uninformative or redundant variables should simplify the learning task without eliminating important predictive information. The observed results provide partial support for this principle but also show its limits. For XGBoost, LGBM, and RF, small SHAP-selected subsets retained strong discrimination. XGBoost achieved its maximum AUC using only 50 SHAP-ranked variables, and LGBM achieved its maximum with 25. In these models, aggressive feature reduction was compatible with performance at least comparable to that of the complete feature set.

GNB showed the clearest counterexample. Its strongest AUC occurred with 25 ANOVA-selected variables, whereas SHAP-based reduction caused pronounced deterioration. This difference is important because it demonstrates that a feature set judged influential within one modelling framework does not automatically preserve the information required by another. MI and ANOVA rank variables independently of the downstream classifier, while SHAP-derived selection reflects how a fitted model uses its variables [22]. A model-dependent importance ranking can therefore encode interactions or dependencies that are useful for the model from which the ranking was obtained but less suitable for another algorithm.

The same principle appears in previous AD feature-selection research. Sparse approaches have demonstrated that restricted predictor sets can support useful AD prediction [13], while feature-selection studies have emphasized that the relative performance of different reduction strategies depends on the data representation and classifier [38]. The present results extend that issue by showing that compression should be regarded as a model-specific intervention rather than a universally transferable preprocessing step.

AUC alone conceals important model behavior

AUC remained an important discrimination measure in the experiment, but the comparison between XGBoost and LR illustrates why it cannot provide a complete account of model behavior.

XGBoost with 50 SHAP variables produced the highest AUC, 0.755, but its balanced accuracy was 0.616. LR with the same feature count and selection strategy achieved a lower AUC of 0.727 but a substantially higher balanced accuracy of 0.668. The two metrics therefore favored different aspects of performance.

This divergence is particularly important in an imbalanced dataset. AUC summarizes a model's ability to rank positive cases above negative cases across classification thresholds. Balanced accuracy instead assigns equal importance to performance in the positive and negative classes [37]. A model may therefore demonstrate strong ranking discrimination yet perform poorly for one class at the operational classification threshold.

The learning curves add another layer. XGBoost's high test AUC coexisted with a persistent train-test performance gap, whereas LR demonstrated a stronger tendency toward convergence as additional training data

were introduced. Consequently, the highest observed AUC did not correspond automatically to the strongest evidence of generalization stability.

This does not invalidate the use of AUC. It shows that feature-compression studies should avoid interpreting a small difference in AUC as sufficient evidence that one compact model is universally preferable. Discrimination, class balance, learning behavior, and error distribution address different aspects of predictive performance.

Compact models create different clinical error trade-offs

The confusion matrices made the distinction between the two leading compact configurations particularly clear. XGBoost missed 289 of the 381 AD cases but generated only 17 false-positive classifications among 2,002 controls. LR missed 134 AD cases but falsely classified 625 controls as AD.

These errors cannot be treated as interchangeable. In a hypothetical risk-stratification setting, missing an individual who later develops AD and incorrectly flagging an individual who does not develop AD can lead to different downstream consequences. The current experiment was not designed to establish which balance is clinically preferable, and such a judgment would require explicit consideration of intended use, prediction thresholds, follow-up procedures, costs, and potential harms.

The appropriate inference is narrower: aggregate metrics conceal materially different error structures. Model selection for early AD risk prediction should therefore include explicit examination of false negatives and false positives rather than relying solely on the highest AUC or a single summary score.

Previous work using EHR-derived data has likewise emphasized the importance of translating statistical prediction into interpretable risk stratification [12]. Explainable AD-prediction models have also sought to make the contribution of individual variables more transparent [39]. The present results indicate that interpretability should extend beyond explaining which features matter to explaining what types of errors a model systematically produces.

Why model-specific feature selection matters

The interaction between feature-selection method and classifier was one of the strongest recurring patterns in the results. SHAP produced excellent compact configurations for several ensemble models and a substantial balanced-accuracy improvement for LR, yet

it performed poorly with GNB. ANOVA produced the best AUC for both GNB and MLP, while MI generally preserved moderate performance without dominating the best results.

These differences can be understood in relation to how selection procedures represent information. ANOVA evaluates univariate class separation. MI evaluates dependence between a feature and the target. SHAP measures contributions in the context of a fitted prediction model. None of these quantities is identical to a universally valid measure of “feature quality.”

SHAP was originally developed as a framework for attributing predictions to input variables [22], and its application to AD and disease-prediction models has increasingly been used to improve interpretability [15]. Using SHAP-selected variables as a reduced input space extends that interpretability logic into feature compression. The present experiment nevertheless shows why such use should remain model-sensitive. Features that are highly influential within a tree ensemble may not provide an equivalent representation for an algorithm whose assumptions and decision structure differ substantially.

Accordingly, the analysis does not support the proposition that SHAP, ANOVA, or MI should be selected in advance as the universally superior approach. The more defensible strategy is to evaluate selection and classifier combinations jointly.

Interpretability and computational implications

Reducing 4,311 candidate variables to 25 or 50 offers an immediate interpretability advantage: the analyst can examine a substantially smaller set of predictors while retaining meaningful discrimination in several configurations. Explainability is particularly relevant in health-related prediction because high-performing models built on thousands of opaque variables may be difficult to interrogate or communicate [40].

The reduced models also provide a more manageable basis for examining whether influential variables are clinically plausible, potentially redundant, or closely related to diagnostic activity. In the current experiment, compact rankings repeatedly brought age, cognitive and memory-related codes, medication patterns, laboratory use, and healthcare-contact variables to the foreground. The source study additionally identifies reduced computational requirements as a practical advantage of feature selection. No direct computational-time or resource measurements were reported, however, so the

present analysis does not quantify that benefit. The defensible conclusion is simply that reducing the number of inputs can simplify the model representation and may reduce processing demands, while the magnitude of any computational advantage was not measured here.

Interpretability must also be distinguished from causality. SHAP values and other feature-importance measures describe how variables contribute to prediction; they do not establish that the selected variables cause AD. The same caution applies to utilization variables and zip code. Their predictive contribution may reflect healthcare access, clinical surveillance, multimorbidity, geographic patterns, or other unmeasured structures. These signals may still be useful for prediction without being appropriate targets for intervention.

Limitations and future directions

Several limitations constrain interpretation of the results. First, the data originated from a single geographic region in Finland. Local characteristics of the population, healthcare organization, coding practices, service availability, and record structures may therefore limit transportability to other regions or health systems. External validation is necessary before assuming that either the compact feature sets or their model-specific performance patterns would reproduce elsewhere.

Second, the final modelling cohort was substantially imbalanced, with 8,005 controls and 1,524 AD cases. Balanced accuracy was included to reduce the interpretive problems associated with majority-class dominance, but the confusion matrices show that class imbalance remained relevant to model behavior. Future work could investigate class weighting, resampling, or other imbalance-management strategies, provided these approaches are evaluated without contaminating independent test data.

Third, the study relied on primary-healthcare and social-service records and did not incorporate specialized biomarker modalities such as neuroimaging. Multimodal studies have shown that imaging and EHR information can be combined in AD prediction [6, 7]. Whether such information would materially alter the feature-compression patterns observed here cannot be determined from the current experiment.

Fourth, hyperparameters were optimized through Grid Search. This provided a systematic search over the defined parameter spaces but did not establish that global optima were found. Alternative optimization procedures were proposed in the source study for future investigation

but were not evaluated experimentally here and should not be interpreted as part of the present results.

Fifth, the learning curves indicated residual overfitting, particularly for XGBoost with 50 SHAP-selected variables. Larger datasets, additional regularization, alternative parameter tuning, and independent external validation could help determine whether this gap reflects limited sample size, model complexity, or dataset-specific structure.

Sixth, feature stability across cross-validation folds was not systematically quantified. A compact feature set is more convincing when the same variables are repeatedly retained under changes in training sample and validation partition. Future analyses should therefore assess selection stability in addition to predictive performance. Finally, feature importance was evaluated primarily as a predictive concept. Several highly ranked variables, including memory-related diagnostic codes and healthcare-use measures, may partially reflect clinical recognition processes occurring before formal diagnosis. The design excluded the final six months before diagnosis, but the extent to which individual features represent underlying disease processes, emerging symptoms, or preceding diagnostic activity remains unresolved.

Conclusion

Reducing a 4,311-variable routine health-record feature space to only 25 or 50 variables can preserve, and in selected model-selection combinations modestly improve, predictive performance for Alzheimer's disease several years before diagnosis. The effect is not universal. It depends strongly on the machine-learning algorithm, feature-selection strategy, and performance dimension being examined.

XGBoost using 50 SHAP-selected variables achieved the highest AUC, whereas Logistic Regression with the same feature count achieved the highest balanced accuracy. Their learning curves and confusion matrices further showed that these aggregate results concealed markedly different behavior: XGBoost retained stronger discrimination but showed greater evidence of overfitting and missed many AD cases, while Logistic Regression detected substantially more AD cases at the cost of considerably more false-positive classifications.

Feature compression should therefore be evaluated as more than a search for the subset producing the highest AUC. Its value depends on how much discrimination is

retained, whether class-balanced performance improves, how stable the model becomes as training data increase, and what kinds of errors the resulting classifier produces. In high-dimensional routine healthcare data, fewer variables can be sufficient—but whether fewer variables produce a better model remains model- and purpose-dependent.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *npj Digit Med.* 2020;3(1). doi:10.1038/S41746-020-0256-0
2. Li Q, Yang X, Xu J, Guo Y, He X, Hu H, et al. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. *Alzheimers Dement.* 2023;19(8):3506-18. doi:10.1002/alz.12967
3. Yang QH, Lyu X, Lin QR, Wang ZW, Tang L, Zhao Y, et al. Effects of a multicomponent intervention to slow mild cognitive impairment progression: a randomized controlled trial. *Int J Nurs Stud.* 2022;125:104110. doi:10.1016/J.IJNURSTU.2021.104110
4. Röhr S, Pabst A, Baber R, Engel C, Glaesmer H, Hinz A, et al. Social determinants and lifestyle factors for brain health: implications for risk reduction of cognitive decline and dementia. *Sci Rep.* 2022;12:12965. doi:10.1038/s41598-022-16771-6
5. Winblad B, Kilander L, Eriksson S, Minthon L, Båtsman S, Wetterholm AL, et al. Donepezil in patients with severe Alzheimer's disease: double-blind, parallel-group, placebo-controlled study. *Lancet.* 2006;367(9516):1057-65. doi:10.1016/S0140-6736(06)68350-5
6. Prabhu SS, Berkebile JA, Rajagopalan N, Yao R, Shi W, Giuste F, et al. Multi-modal deep learning models for Alzheimer's disease prediction using MRI and EHR. *Proceedings of the IEEE 22nd International Conference on Bioinformatics and Bioengineering.* 2022:168-73. doi:10.1109/BIBE55377.2022.00044
7. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci Rep.* 2021;11(1):1-13. doi:10.1038/s41598-020-74399-w
8. Luo H, Hartikainen S, Lin J, Zhou H, Tapiainen V, Tolppanen AM. Predicting Alzheimer's disease from cognitive footprints in mid and late life: how much can register data and machine learning help? *Int J Med Inform.* 2024;190:105540. doi:10.1016/J.IJMEDINF.2024.105540
9. Hossain E, Rana R, Higgins N, Soar J, Datta Barua P, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput Biol Med.* 2023;155:106649. doi:10.1016/J.COMPBIOMED.2023.106649
10. Haq AU, Li JP, Khan J, Memon MH, Nazir S, Ahmad S, et al. Intelligent machine learning approach for effective recognition of diabetes in E-Healthcare using clinical data. *Sensors.* 2020;20(9):2649. doi:10.3390/S20092649
11. Ramkumar G, Seetha J, Priyadarshini R, Gopila M, Saranya G. IoT-based patient monitoring system for predicting heart disease using deep learning. *Measurement.* 2023;218:113235. doi:10.1016/J.MEASUREMENT.2023.113235
12. Tjandra D, Migrino RQ, Giordani B, Wiens J. Cohort discovery and risk stratification for Alzheimer's disease: an electronic health record-based approach. *Alzheimers Dement (N Y).* 2020;6(1). doi:10.1002/TRC2.12035
13. Fukunishi H, Nishiyama M, Luo Y, Kubo M, Kobayashi Y. Alzheimer-type dementia prediction by sparse logistic regression using claim data. *Comput Methods Programs Biomed.* 2020;196:105582. doi:10.1016/J.CMPB.2020.105582
14. Irfan M, Shahrestani S, Elkhodr M. Early detection of Alzheimer's disease using cognitive features: a voting-based ensemble machine learning approach. *IEEE Eng Manag Rev.* 2023;51(1):16-25. doi:10.1109/EMR.2022.3230820
15. Al Olaimat M, Martinez J, Saeed F, Bozdog S. PPAD: a deep learning architecture to predict progression of Alzheimer's disease. *Bioinformatics.*

- 2023;39(Suppl 1):i149. doi:10.1093/BIOINFORMATICS/BTAD249
16. El-Sappagh S, Saleh H, Ali F, Amer E, Abuhmed T. Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time. *Neural Comput Appl.* 2022;34(17):14487-509. doi:10.1007/S00521-022-07263-9
 17. Secondary use of health and social data. Ministry of Social Affairs and Health.
 18. El-Gawady A, Makhlof MA, Tawfik BS, Nassar H. Machine learning framework for the prediction of Alzheimer's disease using gene expression data based on efficient gene selection. *Symmetry.* 2022;14(3):491. doi:10.3390/SYM14030491
 19. Gupta R, Koli N, Mahor N, Tejashri N. Performance analysis of machine learning classifier for predicting chronic kidney disease. 2020 International Conference for Emerging Technology; 2020. doi:10.1109/INCET49848.2020.9154147
 20. Lamba R, Gulati T, Jain A. A hybrid feature selection approach for Parkinson's detection based on mutual information gain and recursive feature elimination. *Arab J Sci Eng.* 2022;47(8):10263-76. doi:10.1007/S13369-021-06544-0
 21. Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient. *Appl Intell.* 2022;52(5):5457-74. doi:10.1007/S10489-021-02524-X
 22. Lundberg SM, Allen PG, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* 2017;30.
 23. Shapley LS. A value for n-person games. *Contributions to the Theory of Games II.* 1953:307-31.
 24. Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. 2018. doi:10.48550/arXiv.1802.03888
 25. Bishop CM. *Pattern Recognition and Machine Learning.* New York: Springer; 2006.
 26. Shabbeer MD, Pravalika M, Leena KG, Likhitha M, Sheshikanth K, Pitchai R. Liver disease prediction model using SVM and logistic regression. *Proceedings of the 2nd International Conference on Edge Computing and Applications;* 2023:403-8. doi:10.1109/ICECAA58104.2023.10212203
 27. Divya R, Shantha Selva Kumari R. Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. *Neural Comput Appl.* 2021;33(14):8435-44. doi:10.1007/S00521-020-05596-X
 28. Arumugam K, Naved M, Shinde PP, Leiva-Chauca O, Huaman-Osorio A, Gonzales-Yanac T. Multiple disease prediction using machine learning algorithms. *Mater Today Proc.* 2023;80:3682-5. doi:10.1016/J.MATPR.2021.07.361
 29. Arslantas MK, Asuroglu T, Arslantas R, Pashazade E, Dincer PC, Altun GT, et al. Using machine learning methods to predict the lactate trend of sepsis patients in the ICU. *Commun Comput Inf Sci.* 2024;2084:3-16. doi:10.1007/978-3-031-59091-7_1
 30. Thotad PN, Bharamagoudar GR, Anami BS. Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods. *Diabetes Metab Syndr.* 2023;17(1):102690. doi:10.1016/J.DSX.2022.102690
 31. Kamarajan C, Ardekani BA, Pandey AK, Chorlian DB, Kinreich S, Pandey G, et al. Random forest classification of alcohol use disorder using fMRI functional connectivity, neuropsychological functioning, and impulsivity measures. *Brain Sci.* 2020;10(2):115. doi:10.3390/BRAINSKI10020115
 32. Mao Y, He Y, Liu L, Chen X. Disease classification based on eye movement features with decision tree and random forest. *Front Neurosci.* 2020;14:547470. doi:10.3389/FNINS.2020.00798
 33. Bikku T. Multi-layered deep learning perceptron approach for health risk prediction. *J Big Data.* 2020;7(1):1-14. doi:10.1186/S40537-020-00316-7
 34. Elujide I, Fashoto SG, Fashoto B, Mbunge E, Folorunso SO, Olamijuwon JO. Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. *Inform Med Unlocked.* 2021;23:100545. doi:10.1016/J.IMU.2021.100545
 35. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min.* 2016:785-94. doi:10.1145/2939672.2939785
 36. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst.* 2017;30:3146-54.
 37. Vanacore A, Pellegrino MS, Ciardiello A. Evaluating classifier predictive performance in multi-class problems with balanced and imbalanced

- data sets. Qual Reliab Eng Int. 2023;39(2):651-69. doi:10.1002/QRE.3237
38. Alshamlan H, Omar S, Aljurayyad R, Alabduljabbar R. Identifying effective feature selection methods for Alzheimer's disease biomarker gene detection using machine learning. Diagnostics. 2023;13(10):1771. doi:10.3390/DIAGNOSTICS13101771
39. Yue L, Chen WG, Liu SC, Chen SB, Xiao SF. An explainable machine learning based prediction model for Alzheimer's disease in China longitudinal aging study. Front Aging Neurosci. 2023;15:1267020. doi:10.3389/fnagi.2023.1267020
40. Biswas AA. A comprehensive review of explainable AI for disease diagnosis. Array. 2024;22:100345. doi:10.1016/j.array.2024.100345