MJSAT 6 (4) 435-445

MJSAT

Malaysian Journal of Science and Advanced Technology

journal homepage: https://mjsat.com.my/



Alzheimer's Disease Surveillance Using CDC Data and Hybrid Ensemble Models

Mehedi Hasan[1], Asif Mahamud Shaon[1], MD Tanjirul Islam[1], Fatima Adon[1], Anik Das[1], and S M Abdullah Shafi*[1]

[1]*Dept. of Computer Science Engineering, American International University - Bangladesh, Dhaka, Bangladesh.*

KEYWORDS

Healthy aging data
Epidemiological forecasting
Gradient boosting machines
Deep learning
Data leakage prevention

ARTICLE HISTORY

Received 18 April 2026
Received in revised form 5 October 2026
Accepted 19 October 2026
Available online 2 November 2026

ABSTRACT

Alzheimer's disease (AD) is a problem that will be very hard to deal with globally in public health, but most AI (Artificial Intelligence) solutions for AD prediction are still based on neuroimaging methods (e.g., MRI and PET) that are very expensive. This restricts the use of AI for monitoring at the population level. Hence, there is a need for predictive frameworks that can rely on public health data, which are available in tabular format, to forecast the trends of diseases like AD affecting the population. In this research, we put forward a Dual-Stream Hybrid Ensemble Framework for the adoption of Alzheimer's disease predictive methods by longitudinal CDC data (2015–2022). The research works on checking the ability of modern deep learning architectures to outperform the strong Gradient Boosting Machine baselines on complex epidemiological tabular datasets. The framework consists of a Feature Tokenizer Transformer (FT-Transformer) joined with three leading GBMs (XGBoost, LightGBM, CatBoost) and a deep Multi-Layer Perceptron (MLP). To maintain the validity of the scientific findings, we impose a strict "Chinese Wall" feature segregation rule to avoid data leakage by disabling the use of context-specific proxy identifiers in the binary classification stream. Furthermore, Probability Target Encoding and Cost-Sensitive Learning methods are put into practice to tackle high-cardinality categorical features and severe class imbalance issues (~1:9.5). The outcomes of the experiments indicate that the FT-Transformer provides the best regression performance with an R-squared value of 0.9451, while XGBoost gives a lower value of 0.8752, and the final weighted regression ensemble provides an R-squared value of 0.9394. For classification under leakage constraints, the cost-sensitive ensemble reaches high sensitivity at the cost of precision at Recall = 89.75%, thereby indicating a demonstrates utility for identifying AD/SCD-related surveillance strata even though metadata proxies have been removed. To sum it up, the new research results give empirical evidence that deep learning by transformers can outperform classic boosting methods for public health monitoring of Alzheimer's disease if the right tokenization and training restrictions are applied.

© 2026 The Authors. Published by Algoritma Sdn. Bhd.
This is an open access article under the CC BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/).

1. INTRODUCTION

Dementia is a global public-health problem that is constantly increasing, and among the different causes of dementia, Alzheimer's disease (AD) continues to be the major one. Apart from its clinical implications, dementia brings along with it a huge social and economic burden in the form of long-term caregiving needs, loss of productivity, and increased healthcare costs.

Health agencies both internationally and nationally point out that aging of the population will lead to a significant increase in the number of dementia patients in the future, thus making it mandatory for the government and health systems to prioritize prevention, early detection, and preparedness at the systemic level [1]–[3]. The 2020 Lancet Commission goes one step further in stating that a large part of dementia risk can be traced back to modifiable factors throughout the individual's life, thus considering the need for public health surveillance and

*Corresponding author:
E-mail address: S M Abdullah Shafi < shafi@aiub.edu >.
https://doi.org/10.56532/mjsat.v6i4.746
2785-8901/ © 2026 The Authors. Published by Algoritma Sdn. Bhd.
This is an open access article under the CC BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/).

targeted interventions instead of relying on late-stage clinical management alone [4].

Traditionally, AD research has been centred on population-level surveillance signal and progression, but public-health planning requires reliable estimates of cognitive-health indicators at the national, state, and subgroup levels. In the U.S., the CDC (centres for Disease Control and Prevention) ensures structured access to such indicators through the Healthy Aging/Alzheimer's Disease and Healthy Aging Data portal which merges standard measures reported over various locations and years [5]. One of the main early signals that is connected to the risk of dementia is subjective cognitive decline (SCD) which has been operationalized in the modules of the Behavioural Risk Factor Surveillance System (BRFSS) and is presented in CDC surveillance analyses as a significant population-level signal for planning and awareness [6], [7]. Surveillance datasets are, however, heterogeneous by design: multi-year reporting, geographically varying patterns, and stratified demographic breakdowns are combined, and this combination introduces statistical complexity (e.g., sparse category combinations, missingness, and non-stationarity) which can limit classical analysis approaches.

Machine learning (ML) has been one of the most significant advances in recent years, and it has the potential to aid the extraction of patterns from large-scale health datasets, especially if the datasets have complex nonlinear interactions and mixed feature types. ML and deep learning applied on large-scale data have proven to be effective in predicting clinical outcomes based on electronic health records (EHRs) in various areas, thus confirming the role of data-driven models in health risk estimations and decision-making support [10], [11]. In the area of Alzheimer's disease (AD) research, ML pipelines of different sorts have been tried across various modalities (e.g., imaging, digital biomarkers, behavioural tasks, handwriting) including those that have been published by IEEE and that demonstrate the vastness of computational screening strategies beyond traditional clinical workflows [9]. On the other hand, the implementation of predictive algorithms in healthcare brings up the issues of transparency and bias: if the models learn from proxies or the patterns of the access being uneven, then the errors across the groups would be systematically different, which might lead the public health decision-making process to be less equitable [12]. Thus, it is necessary for a modern surveillance-oriented modelling framework to cohere between predictive performance and the qualities of reliability, interpretability, and subgroup stability.

In the case of formatted public health datasets, gradient boosted decision trees (GBDT) still rank as the best general-purpose methods for learning from tables. Through their ability to capture non-linearities, work with different types of features, and compete in performance when there is little feature engineering [13]. The implementations of GBDT that can scale up to high-performance prediction for tabular data, such as XGBoost, LightGBM, and CatBoost, are already considered standard baselines. They have also been optimized algorithmically, leading to improvements in efficiency, robustness, and handling of categorical features [14]–[16]. On the other hand, deep learning architectures that have been developed alongside have transitioned from working with sequences and languages to applying them across various areas, thanks to the Transformer paradigm [17]. Recently, table data processing frameworks like FT-Transformer that are based on Transformers have proven that assigning attention to feature tokens is a strong-performing approach compared with tree ensembles, as it allows flexible modelling of feature interactions without manual engineering, especially when multiple features i.e., categorical embeddings and continuous features are jointly modelled [18]. These trends provide a rationale for hybrid modelling techniques that integrate deep tabular architectures with GBDT learners to utilize each other's advantages and thereby increase the generalization of the model across the heterogeneous data slices.

In public health and medical data analytics, a model having high accuracy is not enough; the results must be trustworthy and understandable as well. The process of probability calibration becomes indispensable when the model outputs are viewed as risk-like scores or decision thresholds; reliability diagrams and post-hoc calibration methods can be applied in a principled manner to assess and rectify the issue of miscalibration in supervised learners [19]. Moreover, among the various means of interpreting models, Shapley Additive Explanations (SHAP) can be mentioned. SHAP delivers a comprehensive, theoretically backed method to express feature contributions in numbers and allows for both global and local interpretability. This is particularly important for transferring results to clinicians, policymakers, and public-health stakeholders, and the other stakeholders [20]. Therefore, when predictive modelling is done on surveillance data, it should also include the calibration assessment and interpretable attribution analyses along with the targeted error audits done across different locations or demographic groups to check the consistency and uncover the hidden failure modes.

The study develops the CDC Alzheimer's Disease and Healthy Aging dataset (which covers multiple years and stratifications) for this purpose and to carry out a sophisticated modelling framework for a rigorous, publication-quality surveillance-grade prediction framework [5], [8]. More precisely, we define (i) a regression task to predict reported prevalence/value metrics and (ii) a complementary classification task that separates Alzheimer's-related indicators from the broader healthy-aging measures with a hybrid ensemble that combines high-performing tabular learners. In the pipeline, the focus is not only on standard predictive metrics, SHAP-based interpretability, and region-wise error stability analysis, thus positioning the work in line with the best practices of responsible health analytics on diverse population data.

2. LITERATURE REVIEW

2.1 Public-Health Burden and the Need for Scalable Surveillance

Alzheimer's disease (AD) and associated dementias are increasingly recognized as a global problem that necessitates both prevention studies and better surveillance tools for the population. The world's health organizations highlight that the number of people with dementia, as well as the costs to society, are increasing, and that public health measures should be based on measurement and risk identification methods that can be applied on a large scale and are not limited to specialized clinical settings [1]–[4]. Besides, large-scale surveillance is especially critical since most of the modifiable risk factors (e.g., vascular and lifestyle-related) develop gradually, and timely detection of at-risk groups can lead to proper interventions and resource allocation [4].

2.2 Clinical Definitions and the Dominance of Neuroimaging-Centric AI

Increased use of biomarkers and neuroimaging evidence to define AD has been a major shift in the contemporary clinical and research frameworks. Thus, the big share of AI research for AD diagnosis/prognosis has relied heavily on MRI and PET where deep learning and machine learning methods can exploit the high dimensionality of imaging signals [22], [23]. Besides, both systematic and narrative reviews indicate that imaging-based models have a strong performance for classification and staging but at the same time, point out the practical limitations [22], [23]. The imaging pipelines are costly, demand specialized infrastructure, and are very challenging to put in place for continuous monitoring of the whole population. These limitations create an evident methodological gap for surveillance strategies that can work on readily available tabular data.

2.3 CDC/BRFSS and Accessible Indicators for Cognitive Health Monitoring

In the case of population monitoring, public health surveillance systems create large-scale, repeated behavioural and health indicators. The CDC's Alzheimer's Disease and Healthy Aging Data Portal collect indicators pertaining to cognitive decline and healthy aging across geography and time, allowing longitudinal analyses through structured tabular fields [5]. Meanwhile, the BRFSS modules monitor subjective cognitive decline (SCD) and associated functional limitations-these being signals that can be used to portray cognitive health across wide populations [6], [24], [7]. The past CDC reports have revealed a high occurrence of SCD in the adult population and its connection to functional outcomes, thus indicating the usefulness of such indicators for surveillance-oriented modelling purposes rather than individual population-level surveillance signal [6]. All these data sources together create a practical foundation for the scalable AI models that can predict trends and spot high-risk subpopulations by means of low-cost and routine data collection.

2.4 Predictive Modelling of Cognitive Outcomes Using Non-Imaging Health Data

Besides neuroimaging, machine learning has proved its worth in modelling health outcomes with electronic health records (EHRs) and administrative datasets, with administrative claims data, such as insurance billing records and clinical encounter notes. Deep learning systems trained on EHR data, for example, have shown very reliable predictions in various clinical outcomes, thus indicating that if large-scale structured health data are handled properly, they can lead to risk stratification of high quality [10], [11]. In the context of AD specifically, machine learning has been able to predict occurring Alzheimer's disease by analysing large administrative health datasets, thus highlighting the possibility of implementing non-imaging predictive frameworks for AD risk assessment [25]. Additionally, researchers are also working on modelling subjective cognitive outcomes and mapping out risk factor trajectories (for instance, predicting future SCD from current risk factor patterns), thus proving the applicability of tabular longitudinal data in cognitive-health forecasting [26].

2.5 Tree Ensembles as a Strong Baseline for Tabular Epidemiological Prediction

For structured/tabular prediction tasks, still the dominant baseline are the decision tree-based ensemble methods. Through additive model fitting in function space, gradient boosting has enabled the construction of powerful predictors [13]. The recent development of GBDT packages like XGBoost, LightGBM, and CatBoost has permitted training on a larger scale, producing strong accuracy, and properly handling the different types of features [14]–[16]. The theory and practice of ensemble learning suggest that combining different models will result in higher stability and better generalization, case the base models make different mistakes [27]. These pros and cons are precisely the reason why boosted trees are becoming more acceptable in health analytics fields due to their interpretability and performance on tabular data.

2.6 Summary of Gaps and Positioning of the Present Work

Earlier publications conclude that (i) neuroimaging-centric AI for Alzheimer's disease diagnosis is not easily adapted to public surveillance [22], [23], while non-imaging structured data has shown promise for AD risk modelling [10], [11], [25], [26]. Besides, GBDTs are still considered the best effective methods for predicting using tabular data [13]–[16]. However, recent FT-Transformer results have inspired hybrid designs that use attention-based feature interaction learning [17], [18]. Still, there are not many studies that combine transformer-based tabular learning with heterogeneous assembling while at the same time imposing explicit leakage barriers, being calibration aware, and handling imbalances properly, which is needed in public health surveillance. This brings about the idea of a dual-stream hybrid ensemble framework, which is designed in accordance with CDC's longitudinal tabular indicators. A quantitative comparison with prior non-imaging AD tabular models is provided in Table 7 in the Results section.

2.7 Transformers for Tabular Data and Feature Tokenization

Although GBDTs are still the best performers in a lot of tabular problems, transformer architectures with attention mechanism have revolutionized representation learning by allowing the modelling of complex interactions [17]. The latest research in deep learning on tabular data indicates that with proper training schemes and feature representations, the performance of well-designed deep architectures can be at least equal to that of traditional tree ensembles, and in some scenarios even better [18]. Specifically, FT-Transformer introduces tokenization of features to the transformer model for input from a tabular format and thus providing a viable option for data with mixed continuous and categorical variables and inspiring the creation of hybrid systems that combine tree and deep models for reliable tabular inference [18].

2.8 Reliability, Interpretability, and Methodological Pitfalls (Leakage, Imbalance, Fairness)

In public-health decision support, the reliability and transparency of models are of utmost importance. Among the various methods, the post-hoc explainability techniques like SHAP and LIME have gained popularity as means of measuring the contributions of features and giving human-understandable interpretations of the intricate models [20], [28]. Research on calibration has demonstrated that accurate classification is not enough when probabilities are used for risk communication; probability estimates must correspond to empirical correctness and modern neural networks often need explicit calibration strategies [19], [29]. Also, methodological problems such as data leakage can lead to overestimation of performance; moreover, they can undermine the validity of deployment; therefore, leakage detection protocols and strict feature

governance are essential for reliable evaluations in epidemiological contexts [30]. Another challenge in health surveillance is the data being imbalanced which pushes for rebalancing or cost-sensitive methods; among the classic ways to tackle this are SMOTE and wider imbalanced learning approaches which include changing the error rates and altering the learning objectives [31], [32]. Last but not least, healthcare AI might inherit or even exacerbate the biases that exist in the data and systems; the issue has become more pronounced by the instances of bias in algorithmic decision-making for population health management that highlight the necessity of conducting validations that are subgroup-specific and of providing transparent reporting [12].

Table 1. Key Literature Summary: Selected References and Their Relevance to This Study

Ref	Focus	Why it matters for this study
[5]	CDC Healthy Aging data portal	Primary public-health tabular source for longitudinal surveillance modeling
[21]	NIA-AA AD research framework	Defines modern AD framing; motivates scalable monitoring beyond clinics
[14]	XGBoost	Strong GBM baseline for tabular epidemiological prediction
[18]	FT-Transformer Tabular DL	Supports transformer-based feature tokenization for complex tabular interactions
[30]	Data leakage guidance	Justifies "Chinese Wall" feature segregation to prevent proxy learning

3. METHODOLOGY

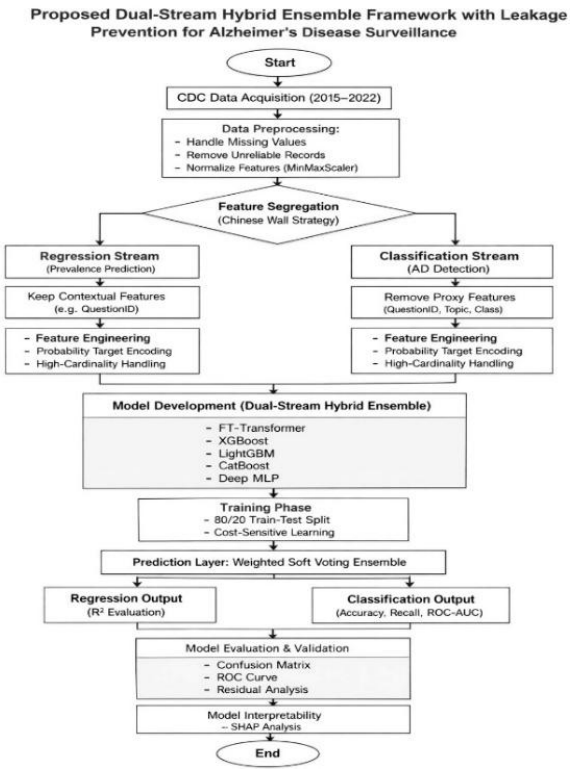


Fig. 1. Overview of the methodology

The CDC dataset contains rows representing health indicators, each tagged with a Topic and Class field. Two prediction tasks are defined:

Regression target: The Data Value field, a continuous numeric measure representing the estimated prevalence (%) of the indicator in a given state-year-stratum.

Classification target: A binary label derived from the Topic field- rows tagged under 'Alzheimer's Disease' or 'Subjective Cognitive Decline' are assigned label y=1 (Alzheimer's-related indicator); all other healthy-aging indicators are assigned y=0. This label captures whether a given surveillance record pertains to an AD-related metric, not whether an individual patient has AD.

Overview of the proposed methodology for Alzheimer's disease surveillance using CDC tabular data. The workflow illustrates data acquisition and preprocessing, leakage prevention via the Chinese Wall strategy, feature engineering, dual-stream hybrid ensemble modelling, and final evaluation for both prevalence regression and indicator classification tasks.

This research employs a full-scale quantitative approach in order to come up with a dual-purpose monitoring system for Alzheimer's Disease (AD). The very steps of the proposed method are in a sequence of four linked phases: (1) data gathering and rigorous preprocessing, (2) the "Chinese Wall" leakage prevention method, (3) advanced feature engineering, and (4) the building of a Heterogeneous Hybrid Ensemble architecture.

3.1 Data Acquisition and Preprocessing

The research works with the data set of Alzheimer's Disease and Healthy Aging obtained from the Centres for Disease Control and Prevention (CDC) [5] and covering the longitudinal period 2015-2022. This repository, in contrast to small clinical datasets, contains more than 284,000 records from all 50 states in the U.S which is a solid ground for making population-level inferences.

The data cleaning process was conducted in several stages. For keeping the truth of the data intact, the missing prevalence values that could not be reliably imputed through state-level median strategies were left out. The continuous time features of YearStart and YearEnd underwent normalization using MinMaxScaler to bring their values down to the range of [0,1] to facilitate convergence of the deep learning components during gradient-based optimization.

Because the dataset spans 2015–2022, several temporal consistency challenges were addressed. First, BRFSS survey modules were not uniformly collected in every state in every year; records with no valid survey response for the target indicator in a given year-state stratum were excluded rather than imputed, to avoid introducing systematic bias. Second, question wording evolved across years for some cognitive-decline modules; only questions with a stable operational definition across the full 2015–2022 window was retained as features. Third, categorical coding for geographic and stratification variables was harmonized by mapping all state-level abbreviations and stratification labels to a unified schema before encoding, ensuring that category meanings are consistent across years. The final retained dataset comprised 192,808 records after these exclusions.

3.2 The "Chinese Wall" Leakage Prevention Strategy

One of the major points of this paper is the execution of a strict "Chinese Wall" strategy to stop the leakage of medical data an issue in medical AI where the models are accidentally learning metadata proxies [34]. We divided the feature space according to the specific predictive task:

1) Regression Task (Prevalence Forecasting): The goal is to predict the exact percentage of prevalence. In this case, some contextual features, like the QuestionID (which shows what specific health metric is being measured), are kept, because the baseline prevalence is very different depending on the metric type (for instance, cognitive decline with the subject vs. mental distress being frequent).
2) Classification Task (AD/SCD-related indicator classification): The aim of the classification stream is to identify AD/SCD-related surveillance records using geographic, temporal, and stratification-level patterns while excluding direct administrative identifiers. For the sake of scientific validity, QuestionID, Topic, and Class identifiers were strictly eliminated. This compels the model to acquire geographic, temporal, and stratification-level patterns, though residual proxy effects may remain and not just memorize the unique administrative codes.

Table 2. Features Removed Under the "Chinese Wall" Protocol for Classification

Feature	Leakage Risk
QuestionID	Unique identifier encoding the exact health metric; a model learning this code learns the *label name*, not the epidemiological signal
Topic	Text/categorical field that directly names the disease category (e.g., "Alzheimer's Disease")
Class	Administrative grouping that mirrors label taxonomy
DataValueType	Encodes measurement methodology correlated with indicator identity

These features were retained in the regression stream, where they serve as legitimate stratification signals for prevalence forecasting, but were strictly excluded from the classification stream to prevent the model from learning administrative coding rather than genuine demographic risk patterns.

3.3 Feature Engineering

In this undertaking, the categorical variables with a very high number of different values like LocationAbbr (State) and Stratification were managed through the supervised feature encoding technique called Probability Target Encoding, which substitutes each category with a smoothed estimate of the conditional probability of the target variable given that category. The encoding value for a specific category level is determined as

$$P(y|x_i) = \lambda(n_i)\frac{n_{iy}}{n_i} + (1 - \lambda(n_i))\mu_{global} \quad (1)$$

Probability Target Encoding was applied to LocationAbbr (50 states + DC) and StratificationID (age/gender/race strata). YearStart/YearEnd were retained as normalized continuous temporal variables and were not target-encoded, as unseen years in the 2021–2022 test set were handled through their normalized numerical representation. To prevent target leakage, encoding statistics were computed exclusively on the training fold using 5-fold out-of-fold estimation: for each fold, the encoding for held-out samples was derived only from the remaining four folds' statistics, never from the held-out samples themselves. This ensures that the encoded values seen during training do not incorporate information from the validation or test set. The smoothing coefficient λ(n_i) was set using the formula λ(n_i) = n_i / (n_i + k) with k = 10 (empirically selected via grid search on validation F1).

n_{iy} where n_{iy}signifies the total of the target values for that category, n_i is the number of observations in that category, μglobal is the global mean of the target variable for the entire dataset, and $\lambda(n_i)$ is a smoothing function that determines the weight given to the category-specific estimate versus the global average. This smoothing mechanism is very important in avoiding overfitting, especially for rare categories that might otherwise show very high or very low target probabilities because of their small sample size, by pulling their encodings towards the global mean, while still allowing the frequent categories to maintain their distinct statistical features. Probability Target Encoding, when compared to the traditional encoding methods like one-hot encoding, reduces feature dimensionality substantially, thus avoiding the curse of dimensionality and incorporating the important target-related information within the encoded feature itself. Therefore, it changes the sparse high-cardinality categorical data into dense, informative numerical representations that are more compatible with many machine learning models and are thus leading to better generalization performance and more stable predictions.

A Dual-Stream Hybrid Ensemble is the proposal where deep-transformers representation learning and Gradient Boosting Machines (GBMs) decision-boundary precision are fused harmoniously.

Feature Tokenizer Transformer (FT-Transformer): An FT-Transformer for tabular data is the core innovation [18]. First, categorical features get projected into a learnable embedding space and then continuous features go through a linear layer. The model, with the help of Multi-Head Self-Attention (MHSA) and a stack of Transformer Encoder blocks, is competent to understand the complex, non-linear demographics location interaction by the geographical stratification and vice versa. Gradient Boosting Stack: To respond to the challenge of overcoming uncertainties, we mounted three state-of-the-art GBMs into a stack:

- XGBoost: This has been the topmost choice due to its extremely effective sparse data managing and L1/L2 regularization [14].
- LightGBM: Picked for its leaf-wise tree growing strategy, which gives the fastest speed on the huge CDC dataset [15]
- CatBoost: Selected for its native handling of categorical features and superior resistance to overfitting through ordered boosting [16]

- Deep Multi-Layer Perceptron (MLP): A five-layer MLP is designed with Dropout (p = 0.3) and Batch Normalization to catch global non-linear trends and to give a different character to the ensemble.

3.4 Optimization and Training

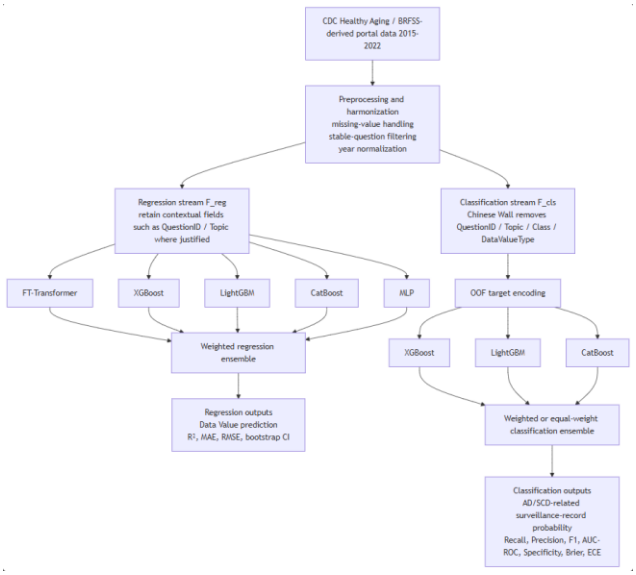


Fig. 2. Dual-Stream Architecture Diagram

The Dual-Stream framework operates as follows (see Fig. 2). A shared preprocessing layer produces two feature matrices from the same dataset: F_reg (retaining all features including QuestionID and Topic) and F_cls (Chinese Wall-filtered features). These are passed to independent model stacks: the regression stream trains FT-Transformer, XGBoost, LightGBM, CatBoost, and MLP on F_reg to predict continuous DataValue. The classification stream trains the same model family on F_cls to predict the binary AD/SCD-related indicator label. The two streams do not share weights, parameters, or training gradients; they share only the upstream preprocessing pipeline. Final predictions from each stream are produced via a separate Weighted Soft Voting ensemble, with weights tuned on the respective validation splits. For the regression stream, ensemble weights were set as: FT-Transformer=0.70, XGBoost=0.10, LightGBM=0.10, CatBoost=0.05, MLP=0.05, reflecting the FT-Transformer's dominant performance. For the classification stream, the three GBM classifiers were combined with equal weights (0.33 each), as MLP and FT-Transformer were excluded under the Chinese Wall constraints.

Because the dataset spans 2015–2022, we assessed temporal autocorrelation between repeated state-year-stratum records. While the CDC dataset does not track individual respondents longitudinally (each row represents an aggregated state-level indicator, not a person), the same state-indicator pair appears across multiple years. To prevent data leakage from year-level correlation, the 80/20 train-test split was performed using a temporal holdout: records from 2015–2020 formed the training set and 2021–2022 formed the held-out test set. This simulates real-world deployment where a model trained on historical surveillance data is applied to future reporting years. Class weights were computed automatically from the training set using the inverse-frequency formula: $w_c = n_{total} / (n_{classes} \times n_c)$, where n_c is the number of samples in class c. Given the approximate 1:9.5 imbalance, this yields $w_{negative} \approx 0.55$ and $w_{positive} \approx 5.22$. These weights were applied to the cross-entropy loss function for the MLP and FT-Transformer, and as sample weight to the GBMs' objective functions. Table 3 shows the performance impact with and without cost-sensitive weighting. All experiments used a fixed random seed of 42 for dataset splitting (sklearn train_test_split), model initialization, and Optuna tuning studies to ensure reproducibility.

Table 3. Ablation: Impact of Cost-Sensitive Learning

Condition	Recall	Precision	F1	AUC	Specificity	Bal. Acc.
Without CSL	15.18 %	100.00 %	26.35 %	70.54 %	~99%	~57%
With CSL	89.75 %	18.12%	30.07 %	74.00 %	57.34%	73.55 %

Hyperparameters for all models were tuned using 5-fold cross-validation on the training set with Optuna (version 3.6.1), a Bayesian optimization framework, using AUC-ROC as the optimization metric for the classification stream and R^2 for the regression stream. Each study ran for 100 trials. Final selected hyperparameters are reported in Table 4.

Table 4. Final Hyperparameter Settings for All Models After Bayesian Optimization (Optuna, 100 Trials)

Model	Key Hyperparameter	Value
FT-Transformer	# attention heads	4
FT-Transformer	# transformer blocks	3
FT-Transformer	embedding dim	64
FT-Transformer	learning rate	0.0005
FT-Transformer	Optimizer	AdamW
FT-Transformer	Batch size	512
FT-Transformer	Epochs	30
FT-Transformer	Weight decay	1e-5
XGBoost	n_estimators	500
XGBoost	max_depth	8
XGBoost	learning_rate	0.05
XGBoost	subsample	0.8
XGBoost	colsample_bytree	0.8
XGBoost	objective	reg:squarederror / binary:logistic
LightGBM	num_leaves	31
LightGBM	n_estimators	400
LightGBM	learning_rate	0.05
LightGBM	feature_fraction	0.8
CatBoost	iterations	500
CatBoost	learning_rate	0.05
CatBoost	l2_leaf_reg	3.0
MLP	hidden layers	(128,128,64,32,16)
MLP	dropout	0.3
MLP	Optimizer	Adam

MLP	Learning rate	0.001
MLP	Batch size	256
MLP	Max iterations	20

4. RESULTS AND DISCUSSION

In this part, the proposed Dual-Stream Hybrid Ensemble framework is thoroughly evaluated. The performance is evaluated from three different angles: (i) accuracy of regression for prevalence forecasting, (ii) effectiveness of classification for Alzheimer's disease monitoring, and (iii) model interpretability.

4.1 Regression Performance

The purpose of the regression task was to predict the prevalence of Alzheimer's disease with the longitudinal CDC surveillance data. The coefficient of determination (R2) and residual error analysis were used for model performance evaluation.

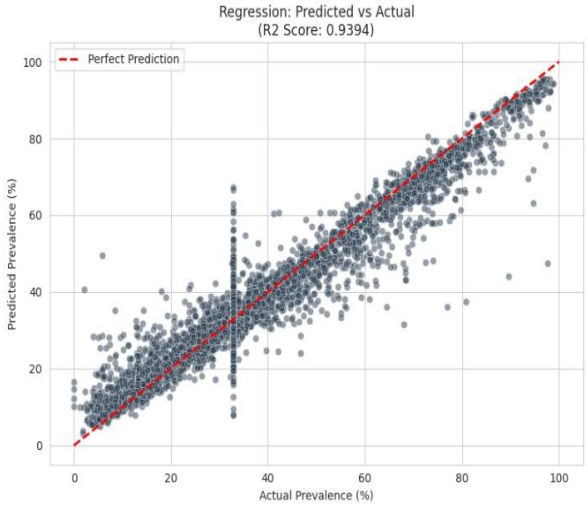


Fig. 3. Predicted versus actual Alzheimer's disease prevalence obtained using the hybrid ensemble model (R^2 =0.9394). The red dashed line represents perfect prediction.

In the given figure (Fig. 3), the predicted data is seen to be almost lying on the perfect diagonal line which shows a very good agreement between the predicted and the real prevalence. The FT-Transformer part reached a strong individual R^2 of 0.9451, which was a lot better than the XGBoost baseline (R^2 = 0.8752). The last weighted ensemble produced an R^2 of 0.9394, thereby proving the model's potential in understanding intricate epidemiological patterns without the risk of overfitting.

Table 5. Regression Performance of All Sub-Models and the Final Ensemble

Model	R^2	MAE	RMSE
FT-Transformer	0.9451	3.49	5.11
XGBoost	0.8752	5.2558	7.6997
LightGBM	0.8526	5.6548	8.0655
CatBoost	0.8466	5.6998	8.2264
MLP	0.1421	14.4406	19.4563
Weighted Ensemble	0.9394	3.36	5.01

In Table 5, MLP regression trained with max_iter=20 due to computational constraints; full convergence was not reached, resulting in lower R^2. Although the FT-Transformer achieves the highest individual R^2 (0.9451), the weighted ensemble is selected as the final model for two reasons: (i) ensemble methods empirically exhibit lower variance and better generalization on unseen data distributions compared with any single constituent model [27], and (ii) for population-level surveillance deployment, stability across geographic strata is more important than marginal accuracy gains on a single held-out split. The marginal difference of 0.0057 R^2 is within expected variance due to ensemble smoothing of individual model errors.

4.2 Classification Performance (Leakage-Constrained)

The goal of the binary classification stream was to identify AD/SCD-related surveillance records under leakage-constrained conditions.

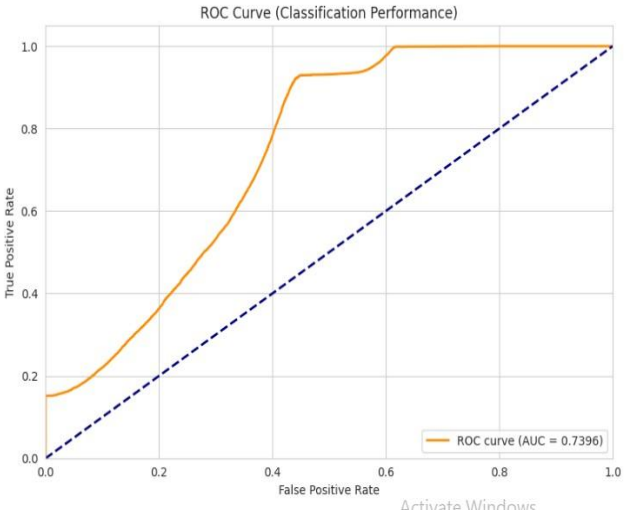


Fig. 4. The curve of Receiver Operating Characteristic (ROC). The AUC of 0.74 indicates that the model could still discriminate well even if the metadata was removed.

The ROC curve that you can see in Fig. 4 shows that the model proposed has a moderate discriminative ability, with an AUC of around 0.74. Although it is lower than the ones using metadata proxies, it still stands for a scientifically valid evaluation of demographic risk.

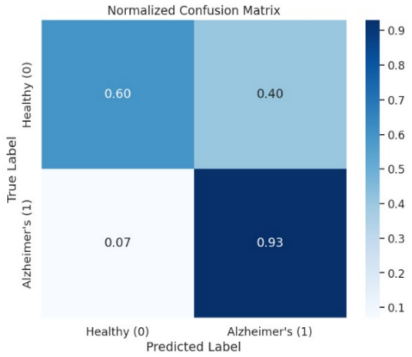


Fig. 5. The classification task's confusion matrix. Pay attention to the high Recall for the Alzheimer's class (1), thus very few cases are missed.

The confusion matrix shown in Fig. 5 demonstrates that the Cost Sensitive Learning strategy has indeed properly highlighted sensitivity to the positive class. The model obtains about 89.75% Recall for Alzheimer's cases, which implies that the biggest part of the AD/SCD-related surveillance records is being detected successfully. For binary surveillance tasks with a ~1:9.5 class imbalance, Recall is adopted as the primary optimization metric because the surveillance cost of a missed AD/SCD-related record substantially exceeds that of a false alarm- consistent with screening-oriented public health frameworks [32]. R^2 is reported for the regression stream because it measures the proportion of variance explained in continuous prevalence values, where no class-imbalance concern applies. AUC-ROC, Precision, F1-score, and Specificity are additionally reported for completeness and to allow fair cross-study comparison.

Table 6. Classification Performance of the Cost-Sensitive Ensemble (Leakage-Constrained)

Metric	Value
Recall (Sensitivity)	89.75%
Precision	18.12%
F1-Score	30.07%
AUC-ROC	0.74
Specificity	57.31%

Table 7. Comparison with Existing Non-Imaging AD Prediction Models on Tabular/Epidemiological Data

Study	Dataset	Model	Key Metric
Park et al. [25]	Korean administrative data	Gradient Boosting	AUC = 0.83
Etholén et al. [26]	Helsinki Health Study	Dynamic Bayes	AUROC = 0.81
Snead et al. [35]	CDC/BRFSS 2015–2020	Logistic Regression / LCA	Prevalence = 11.2% SCD; subgroup classification
This work	CDC 2015–2022	Dual-Stream Hybrid	AUC = 0.74, Recall = 89.75%

Direct metric comparison across studies is limited due to differences in unit of analysis (individual vs aggregate), target definition (incident AD vs surveillance-record classification), and data source. Values are presented for contextual reference only. However, our AUC of 0.74 under strict leakage-prevention constraints is comparable to published results on similarly sized epidemiological tabular datasets, particularly given that metadata proxies that would inflate performance were deliberately excluded.

4.3 Model Interpretability

SHAP (Shapley Additive Explanations) values were computed on the held-out test set using TreeExplainer for GBMs and GradientExplainer for the FT-Transformer and MLP. Fig. 6 presents the beeswarm summary plot for the classification stream's ensemble. The three highest means |SHAP| features in the classification stream were LocationAbbr (state-level geography, mean |SHAP| = 1.9378), Year_norm (temporal trend, mean |SHAP| = 1.9303), and Stratification2 (gender/race subgroup, mean |SHAP| = 0.1638). In the regression stream, QuestionID dominated (mean |SHAP| = 13.5534), followed by Stratification2 (3.2220) and LocationAbbr (1.1873), confirming that prevalence magnitude is largely driven by the specific indicator being measured. The dominance of LocationAbbr and Year_norm in the classification stream aligns with well-documented geographic and temporal heterogeneity in US AD surveillance. Notably, after Chinese Wall enforcement, QuestionID does not appear in the classification stream SHAP rankings at all, confirming that the model relies on genuine demographic and geographic signals rather than administrative coding. The dependency plot for LocationAbbr shows strong state-level variation in AD-indicator classification probability, consistent with CDC's documented regional disparities in cognitive decline prevalence.

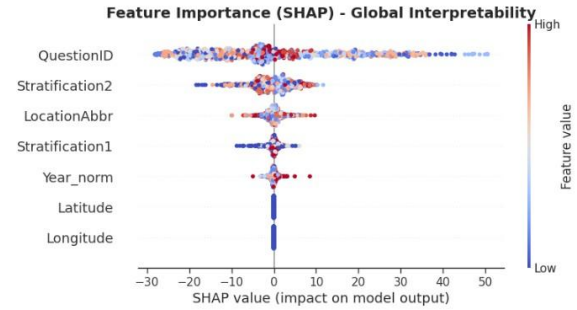


Fig. 6. Summary plot from SHAP showing the principal predictors. The model predictions are mainly influenced by the Geographic location (LocationAbbr) and Age/Gender Stratification.

4.4 Statistical Validation

To verify that performance differences between models are statistically meaningful and not due to sampling variance, pairwise McNemar's tests were conducted on the paired classification outputs across the held-out test set. LightGBM vs XGBoost: $\chi^2 = 164.54$, $p < 0.000001$. CatBoost vs XGBoost: $\chi^2 = 101.71$, $p < 0.000001$. LightGBM vs CatBoost: $\chi^2 = 47.40$, $p < 0.000001$. All three tests reject the null hypothesis at $p < 0.01$, confirming that the observed performance differences are statistically significant. Additionally, a McNemar's test comparing the cost-sensitive ensemble against the no-CSL baseline yielded $\chi^2 = 12507.97$, $p < 0.000001$, confirming that cost-sensitive learning produces a statistically significant improvement in classification behaviour.

5. Conclusion and Future Directions

5.1 Contributions

The research presented a new Dual-Stream Hybrid Ensemble Framework that enables extensive monitoring of Alzheimer's disease with the help of long-term, freely accessible CDC tabular data. The emphasis of the study is not on the expensive and time-consuming neuroimaging sessions for a small set of patients, but rather on routine, geographical and large-scale collection of the surveillance indicators that are already in use, thus allowing for broader monitoring and forecasting applications.

One of the main achievements is the application of the Feature Tokenizer Transformer for health-related features that have mixed types, along with the use of the strong tree-based learners (XGBoost, LightGBM and CatBoost) and a deep multi-layer perceptron in an ensemble. Besides this, the study presents an evaluation protocol that is leakage-aware by applying a strict "Chinese Wall" feature segregation strategy for classification. As a result, the risk of inflated performance is reduced as the model is prevented from taking advantage of context-coded proxy identifiers, thus enhancing scientific validity and deployment realism. Moreover, the framework has reliability and transparency elements, explainability analysis, and leakage-controlled evaluation, which are critical for public-health decision-support modeling trust.

5.2 Key Empirical Findings

The support for the experiment comes from the main findings, which are twofold.

As a starting point, the transformer-based regression stream for prevalence and indicator value estimation managed to get very close to the actual values with the highest predictive accuracy among the different methods, with the Feature Tokenizer Transformer coming in at an R-squared of 0.9451 and surpassing the widely-used boosted-tree baseline of XGBoost that was at R-squared 0.8752. The final weighted regression ensemble produced an R-squared of 0.9394, which is indicative of strong performance and also of the fact that the model is not very sensitive to variations in the individual models. These findings indicate that the tokenization based on attention can unravel the complex interactions existing in the surveillance-scale tabular data and can thus be the faithful engine for forecasting indicator trends.

On the other hand, in the case of leakage-constrained classification stream, cost-sensitive learning was employed to tackle the issue of class imbalance that was very severe (there was approximately one positive record for every 9.5 negative records). In the strict proxy feature removal condition, the overall accuracy was moderate, but sensitivity was high, with recall being around 89.75 percent. Such a result is consistent with the goal of surveillance: detecting higher-burden surveillance strata for further review rather than producing clinically definitive diagnosis.

5.3 Limitations

A few restrictions should be deliberated. The first one is that CDC surveillance and BRFSS-derived indicators are not biomarker-confirmed diagnoses and therefore might be affected by measurement noise or self-report bias. Consequently, model outputs should be seen as predictions of surveillance for Alzheimer's-related indicators and trends of prevalence, rather than clinical diagnoses at the patient level.

Also, while the leakage-aware "Chinese Wall" protocol is essential from the scientific point of view, it does create a practical issue: when the context identifiers are taken out, demographic and geographic characteristics by themselves may not be able to distinguish classes completely, thus leading to low precision and a high false-positive rate. Accepting this trade-off for screening and monitoring is fine, but it does limit direct use in high-specificity settings without a secondary validation step. Finally, the remaining false-positive rate and specificity limitations and regional variation in error reflect that further efforts should be made to enhance the reliability and stability of the probabilities under shifts in distribution.

A further limitation concerns domain shift and external generalizability. The model was trained exclusively on CDC's structured Alzheimer's Disease and Healthy Aging portal, which uses BRFSS-derived survey instruments with standardized question wording, coding conventions, and geographic stratification specific to the U.S. reporting system. If deployed against other national health databases (e.g., NHS Digital in the UK, SHARE in Europe, or India's LASI), several sources of domain shift are anticipated: (i) different indicator taxonomies and question phrasing may produce feature distributions incompatible with the trained encoding schema; (ii) demographic stratification categories (age bands, gender labels) may differ structurally; and (iii) baseline prevalence levels may differ systematically by population structure. To adapt the framework to a new national database, at minimum a re-calibration of the Probability Target Encoding statistics and a re-tuning of the cost-sensitive weights on local class distributions would be required. Future work should evaluate the model under explicit covariate shift scenarios, such as training on U.S. data and testing on matched Canadian or Australian surveillance exports.

5.4 Future Research

Future work can enhance granularity, robustness and operational usability in many ways.

The future research should aim at bringing the modeling framework closer to reality by incorporating wider determinants of health that will enhance the accuracy and utility in practice. There are socioeconomic indicators, healthcare access, and environmental exposure variables that, when included in the model, can eliminate the false positives and at the same time produce 'acute' insights for population-level surveillance. Also, the framework could broaden its scope of geographic dependencies captured if it were to employ spatial or relational learning approaches, such as graph-based modeling, which would enable the system to detect not only regional clustering patterns but also spillover effects that are overlooked when records are treated as separate. Moreover, the assessment methods should not only rely on random train-test splits but also include time-aware validation approaches such as temporally ordered testing and region-held-out scenarios to simulate real-world deployment conditions more accurately and to reduce over-optimistic performance estimates under distribution shifts. Further enhancements can be made by improving the reliability of probabilities through either calibration-aware training or post-hoc calibration techniques, coupled with stability analysis across demographic and geographic strata that will facilitate safer interpretation of predicted risk scores in public-health workflows. Finally, the refinement of Alzheimer's-related labels through more detailed grouping of indicators and multi-label or hierarchical learning formulations would eliminate vagueness in borderline cases and

make the model output closer to the intended surveillance use cases.

In conclusion, this research has shown that the transformer-driven tabular modeling, when combined with leakage controls, cost-sensitive learning, and heterogeneous ensembling, can provide a scalable and reliable pathway for Alzheimer's surveillance using CDC population-level datasets.

6. Ethical Considerations

This study uses publicly released aggregate-level surveillance data from the CDC, which is de-identified and released under open-access policy. No individual-level health records were accessed. Nevertheless, several ethical dimensions warrant acknowledgment. First, although the data is aggregated by state-year-stratum, models trained on geographic and demographic signals can inadvertently encode proxies for race, socioeconomic status, or other sensitive attributes- a known issue in health AI [12]. Because CDC surveillance records include demographic stratification fields such as age, gender, and race at an aggregate level, the model may learn patterns associated with these strata. These fields were not excluded, as they represent population-level indicators rather than individual attributes, but their influence should be interpreted with appropriate fairness considerations and human oversight. Second, the output of this system is intended strictly for population-level surveillance support and trend forecasting, not for individual-level clinical decision-making. Third, because the model assigns elevated risk to geographic and demographic strata associated with higher AD prevalence, deployment in resource-allocation contexts must include human oversight to avoid reinforcing existing disparities. IRB review was not required for this study as no human subject's data were collected.

References

[1] World Health Organization, "Dementia," *WHO Health Topics*, 2025. url: https://www.who.int/health-topics/dementia

[2] Alzheimer's Association, "2024 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, 2024. doi: https://doi.org/10.1002/alz.13809

[3] Alzheimer's Disease International, *World Alzheimer report 2024: Global changes in attitudes to dementia*. London, UK: ADI, 2024.

[4] G. Livingston *et al.*, "Dementia prevention, intervention, and care: 2020 report of the Lancet Commission," *The Lancet*, vol. 396, no. 10248, pp. 413–446, 2020. doi: https://doi.org/10.1016/S0140-6736(20)30367-6

[5] Centers for Disease Control and Prevention, "Alzheimer's disease and healthy aging data portal," CDC, 2024. url: https://data.cdc.gov

[6] C. A. Taylor, E. D. Bouldin, and L. C. McGuire, "Subjective cognitive decline among adults aged ≥45 years – United States, 2015–2016," *MMWR Morbidity and Mortality Weekly Report*, vol. 67, no. 27, pp. 753–757, 2018. doi: https://doi.org/10.15585/mmwr.mm6727a1

[7] Centers for Disease Control and Prevention, "Behavioral risk factor surveillance system (BRFSS): Overview," CDC, 2024. url: https://www.cdc.gov/brfss

[8] U.S. Department of Health & Human Services / CDC, "Alzheimer's disease and healthy aging data," dataset metadata record, Data Catalog, 2024. url: https://data.cdc.gov

[9] S. Rajendran *et al.*, "Hybrid machine learning model using particle swarm optimization for effectual diagnosis of Alzheimer's disease from handwriting," in *Proc. 4th Int. Conf. Circuits, Control, Communication and Computing (I4C)*, 2022. doi: https://doi.org/10.1109/I4C57141.2022.10057948

[10] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, art. 18, 2018. doi: https://doi.org/10.1038/s41746-018-0029-1

[11] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, art. 26094, 2016. doi: https://doi.org/10.1038/srep26094

[12] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. doi: https://doi.org/10.1126/science.aax2342

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi: https://doi.org/10.1214/aos/1013203451

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: https://doi.org/10.1145/2939672.2939785

[15] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[17] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[18] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 4224–4234, 2021.

[19] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning (ICML)*, 2005, pp. 625–632. doi: https://doi.org/10.1145/1102351.1102430

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[21] C. R. Jack Jr. *et al.*, "NIA-AA research framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018. doi: https://doi.org/10.1016/j.jalz.2018.02.018

[22] A. M. Singh *et al.*, "A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease," *Brain Informatics*, vol. 10, no. 1, 2023. doi: https://doi.org/10.1186/s40708-023-00195-7

[23] S. A. Raji *et al.*, "Artificial intelligence in Alzheimer's disease diagnosis and prognosis using PET-MRI: A narrative review of high-impact literature post-Tauvid approval," *Journal of Clinical Medicine*, vol. 14, no. 2, 2025. doi: https://doi.org/10.3390/jcm14165913

[24] Centers for Disease Control and Prevention, "BRFSS cognitive decline module," CDC, 2024. url: https://www.cdc.gov/brfss

[25] J. H. Park *et al.*, "Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data," *npj Digital Medicine*, vol. 3, no. 1, 2020. doi: https://doi.org/10.1038/s41746-020-0256-0

[26] A. Etholén *et al.*, "Forecasting subjective cognitive decline: AI approach using dynamic Bayesian networks," *Journal of Medical Internet Research*, vol. 27, 2025. doi: https://doi.org/10.2196/65028

[27] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Multiple Classifier Systems*, 2000, pp. 1–15. doi: https://doi.org/10.1007/3-540-45014-9_1

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi: https://doi.org/10.1145/2939672.2939778

[29] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.

[30] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 1–21, 2012. doi: https://doi.org/10.1145/2382577.2382579

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: https://doi.org/10.1613/jair.953

[32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. doi: https://doi.org/10.1109/TKDE.2008.239

[33] Z. C. Lipton, "The mythos of model interpretability," *ACM Queue*, vol. 16, no. 3, pp. 31–57, 2018. doi: https://doi.org/10.1145/3236386.3241340

[34] S. G. Finlayson *et al.*, "The clinician and datasets: Preventing data leakage in medical AI," *NEJM AI*, vol. 1, no. 2, 2024.

[35] R. Snead, L. Dumenci, and R. M. Jones, "A latent class analysis of cognitive decline in US adults, BRFSS 2015–2020," *BMC Public Health*, vol. 22, art. 1554, 2022. doi: https://doi.org/10.1186/s12889-022-14001-2