

Article

Artificial Intelligence–Enabled Dementia Risk Prediction for Smart and Sustainable Healthcare: An Interpretable Machine Learning Study Using NHATS

Ashrafe Alam ^{1,*}, Md Golam Rabbani ¹ and Victor R. Prybutok ²¹ Department of Information Science, University of North Texas, Denton, TX 76203, USA; mdrabbani@my.unt.edu² G. Brint Ryan College of Business, University of North Texas, Denton, TX 76203, USA; victor.prybutok@unt.edu

* Correspondence: ashrafealam@my.unt.edu

Abstract

Dementia is an increasing public health challenge, yet scalable methods for early risk detection using non-clinical data remain limited. This study develops and evaluates interpretable machine learning models to predict dementia risk among older adults using nationally representative longitudinal data. Data were sourced from the National Health and Aging Trends Study (NHATS, 2011–2022) and included 5984 community-dwelling U.S. adults aged 65 and older who were dementia-free at baseline. Dementia onset was identified using the validated NHATS classification algorithm based on cognitive assessments, proxy reports, and physician diagnoses. After data preprocessing and feature engineering, missing values in continuous variables were imputed with k-nearest neighbors, while categorical variables were handled via one-hot encoding and mode-based imputation. Five supervised machine learning algorithms were trained and evaluated through stratified cross-validation, using performance metrics that account for class imbalance. Among these models, XGBoost showed the strongest overall performance, achieving the highest classification accuracy (0.881 ± 0.004), the lowest Brier score (0.094 ± 0.002), and the highest ROC–AUC (0.823 ± 0.005), with RF showing comparable results. Explainable AI analyses with SHapley Additive exPlanations (SHAP) consistently identified digital technology use, outdoor activity frequency, and social network size as the most influential predictors across models. These findings indicate that interpretable machine learning based on non-clinical, modifiable behavioral and social factors can support scalable, prevention-focused dementia risk assessment and inform prevention-oriented strategies that promote digital inclusion and social engagement among older adults.

Keywords: artificial intelligence; interpretable machine learning; explainable AI; dementia risk prediction; National Health and Aging Trends Study (NHATS); digital technology use; social engagement; SHAP; sustainable healthcare

Academic Editors: Martina Iammarino, Chiara Verdone and Stefano Galantucci

Received: 22 January 2026

Revised: 18 February 2026

Accepted: 20 February 2026

Published: 24 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Dementia is an increasing public health challenge worldwide. Over 55 million people currently have dementia globally, and the number is expected to surpass 140 million by 2050 because of aging populations [1,2]. Around 10 million new cases are reported

annually. Without disease-modifying treatments, attention has increasingly turned to early risk detection and prevention strategies to delay onset and reduce the long-term societal burden [3,4].

Traditional dementia risk prediction methods primarily rely on clinical, neuropsychological, and biological markers, such as cognitive test scores, neuroimaging, genetic risk factors, and medical diagnoses [5]. However, such models often achieve high predictive accuracy; they usually rely on expensive, invasive, or clinic-based data, which limits their scalability for population-level screening and public health applications [6,7]. Additionally, several existing models lack transparency, raising concerns about interpretability and trust when used in high-stakes health decisions [8].

Recent advances in ML enable modeling of complex nonlinear relationships among diverse risk factors in large-scale observational data [9]. However, much of the ML literature on dementia prediction continues to focus on clinical or imaging-based features and cross-sectional datasets, with limited attention to modifiable behavioral and social determinants [10,11]. This causes a gap between highly advanced prediction models and prevention-focused, scalable tools for real-world public health use.

Growing epidemiological evidence indicates that non-clinical, adjustable factors such as social engagement, physical activity, and digital technology use play a significant role in cognitive aging and dementia risk [12,13]. Recent large-scale studies and meta-analyses indicate that regular use of digital technologies is associated with slower cognitive decline and a reduced incidence of dementia, supporting the concept of a technological reserve analogous to cognitive reserve [14–16]. These behavioral and social factors are routinely collected in national aging surveys, making them promising candidates for scalable, prevention-focused risk modeling.

At the same time, concerns about the interpretability of complex ML models have increased interest in explainable artificial intelligence (XAI) techniques. Methods such as SHapley Additive exPlanations (SHAP) enable transparent attribution of model predictions to individual features, helping to bridge the gap between predictive performance and actionable insights [11,17]. Combining XAI with longitudinal, non-clinical data offers a promising approach to developing interpretable and policy-relevant dementia risk prediction models.

Against this backdrop, the present study develops and evaluates interpretable ML models for dementia risk prediction using over a decade of nationally representative longitudinal data from the National Health and Aging Trends Study (NHATS). Unlike previous NHATS-based studies that mainly examine associations through regression, this study uses a comparative, calibration-aware machine learning framework with cross-model explainability to assess predictor stability, calibration, and interpretability.

2. Related Work

2.1. Machine Learning–Based Dementia Risk Prediction

Over the past decade, ML techniques have increasingly been used to predict dementia and Alzheimer's disease using algorithms such as logistic regression, support vector machines, random forests, k-nearest neighbors, and gradient boosting. Many recent studies report high classification accuracy, especially when using clinical or neuroimaging datasets. For example, ensemble and boosting methods applied to memory clinic or neuroimaging cohorts have achieved over 90% accuracy in distinguishing Alzheimer's disease or mild cognitive impairment from cognitively normal controls [18,19].

Despite these advancements, several limitations still exist. Many ML models are trained on highly selected clinical samples or balanced research datasets, which raises concerns about overfitting and their limited generalizability to community-dwelling populations [7]. Predictive performance often depends heavily on biomarkers or cognitive test

scores that are impractical for large-scale screening [20]. Additionally, many studies focus on accuracy while giving limited attention to calibration, interpretability, or real-world deployment considerations [21,22].

2.2. Behavioral, Social, and Digital Predictors

Recent research has increasingly explored how lifestyle, behavioral, and social factors are incorporated into dementia risk prediction models. Epidemiological studies consistently show that social isolation, decreased physical activity, and limited cognitive engagement are connected to a higher risk of dementia [4,12]. Several ML-based studies have demonstrated that non-clinical variables can substantially improve predictive accuracy when combined with demographic and health data indicators [23–25].

Digital technology use has emerged as a notably promising predictor of behavior. A large meta-analysis involving over 400,000 older adults revealed that general technology use is associated with significantly lower risks of cognitive impairment and slower cognitive decline [26]. Cohort studies further suggest that discontinuing digital activities is associated with faster memory decline, whereas continued engagement is associated with protective effects [27–29]. While existing NHATS-based studies offer valuable insights into the relationship between technology use and cognitive trajectories, they mostly rely on single-model or regression-based analyses. These studies generally do not evaluate predictor stability, calibration, or cross-model consistency using explainable ML techniques.

2.3. Explainable AI and Longitudinal Cohorts

Model interpretability is increasingly recognized as vital for adopting ML in healthcare, especially for dementia risk assessment [8,30]. Explainable AI techniques, such as SHAP and LIME, have been used in dementia prediction models to identify influential features and verify model behavior against domain standards knowledge [31,32]. However, most explainable ML studies still rely on clinical or imaging-based predictors, which restricts their usefulness for scalable public health applications.

Longitudinal population-based cohorts such as NHATS and the Health and Retirement Study offer opportunities to develop generalizable, prevention-focused dementia risk models using non-clinical data [33,34]. However, comprehensive and interpretable ML frameworks that utilize these datasets remain scarce. This study addresses that gap by combining explainable ML methods with longitudinal, nationally representative data and adaptable behavioral predictors.

2.4. Study Contributions

This study adds to the dementia risk prediction literature by presenting a thoroughly tested, interpretable AI framework based on ML designed for longitudinal, non-clinical population data. Instead of creating a new learning algorithm, the focus is on methodological integration, rigorous evaluation, and practical application of ML in dementia prevention.

The key contributions are as follows:

2.4.1. Integrated Interpretable ML Framework

We propose a unified framework that integrates longitudinal outcome modeling, calibration-aware evaluation, and explainable artificial intelligence. Unlike previous studies that mainly focus on accuracy or cross-sectional prediction, this work jointly evaluates discrimination, performance sensitive to imbalance, and probabilistic calibration through strict stratified cross-validation and leakage control, ensuring reliable real-world risk estimation [35]. This framework builds on previous NHATS-based analyses by shifting from

mere association testing to systematically assessing predictive robustness, calibration, and explainability across various machine learning paradigms [36,37].

2.4.2. Emphasis on Modifiable, Non-Clinical Predictors

Using nationally representative NHATS data spanning over a decade, the study demonstrates how modifiable behavioral and technological factors can be systematically integrated into ML models [38,39]. This data-driven approach supports scalable, prevention-focused dementia risk assessment beyond clinic-based or biomarker-dependent methods [40].

2.4.3. Cross-Model Interpretability and Predictor Stability

By incorporating SHAP-based explanations, regression estimates, and nonparametric statistical tests, the study evaluates the consistency of influential predictors across various algorithms. The convergence of key features across models provides strong evidence for risk signals rather than model-specific effects [41,42].

2.4.4. Applied and Translational Contribution

The study aligns healthcare-focused ML priorities that emphasize transparency, reproducibility, and deployability. The proposed framework connects predictive modeling to actionable insights, supporting dementia prevention on a population level and informing public health decisions. To achieve this, the study addresses the following research questions:

RQ1. Do modifiable behavioral and technological factors improve dementia risk prediction when combined with demographic and health-related variables in a longitudinal setting?

RQ2. How do commonly used ML algorithms compare in predictive performance and calibration within an interpretable framework?

RQ3. Which predictors consistently emerge as influential across models, and how stable are these signals?

RQ4. How effectively can explainable ML methods support transparent and prevention-oriented dementia risk assessment?

3. Materials and Methods

3.1. Dataset Source and Study Sample

We used data from the National Health and Aging Trends Study (NHATS), a nationally representative longitudinal panel survey of Medicare beneficiaries aged 65 years and older in the United States. NHATS began in 2011 (Round 1) with an initial sample of 8245 older adults and has conducted annual follow-up interviews since then, with periodic cohort replenishments to maintain national representativeness. The present analysis includes data from NHATS Rounds 1–12 (2011–2022). Our study focused on community-dwelling participants (i.e., individuals living in private residences or non-nursing home settings) who were dementia-free at baseline. Participants diagnosed with dementia or Alzheimer's disease at baseline were excluded to identify incident dementia cases during follow-up. We also excluded individuals with substantial missing data on key predictors (see Section 3.3) or those who were lost to follow-up before outcome ascertainment. The baseline NHATS cohort included 8245 participants. Individuals with prevalent dementia at baseline were excluded, along with participants with significant missing data on key predictors and those who lost to follow-up before dementia status could be determined. After applying these exclusion criteria, the final analytic sample consisted of 5984 participants (Figure 1).

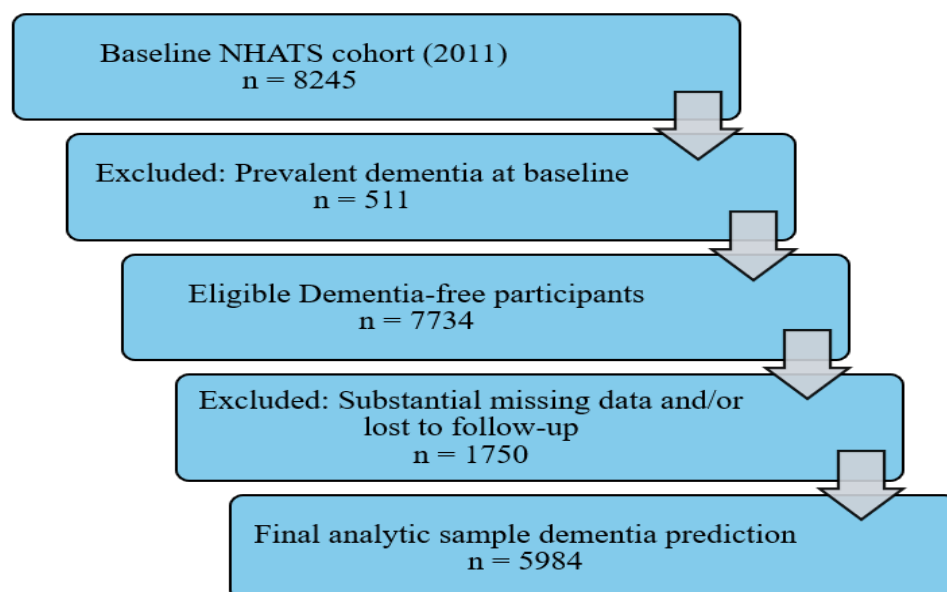


Figure 1. Participant selection for the analytic NHATS cohort (2011–2022). Arrows indicate the flow of participants through each stage of inclusion and exclusion.

Sampling weights and variables that account for the complex survey design were used where appropriate to ensure that all estimates accurately reflect the U.S. older adult population.

3.2. Outcome Variable (Dementia Onset)

The primary outcome was a binary indicator of probable dementia status based on the validated NHATS algorithm [43]. Participants were classified as having dementia if they met at least one of the following criteria: (a) a self- or proxy-reported physician diagnosis of dementia or Alzheimer’s disease; (b) performance ≥ 1.5 standard deviations below the mean in two or more cognitive domains (memory, orientation, executive function); or (c) a proxy-completed AD-8 dementia screening score ≥ 2 , indicating cognitive impairment [44]. Cognitive domains were assessed using immediate/delayed word recall, orientation questions (date and political figures), and a clock-drawing task. For those unable to complete cognitive testing, the AD-8 informant tool was administered. Participants were labeled “1” if they met the criteria in any follow-up wave (2012–2022), and “0” otherwise. This definition is consistent with NHATS protocols and has demonstrated validity against clinical diagnoses.

3.3. Predictor Variables and Feature Engineering

We examined a rich set of predictor variables encompassing demographic, social, technological, behavioral, and clinical domains, selected based on literature linking them to cognitive health [29,45].

Demographics: Age (years), sex (male or female), educational attainment (categorized as less than high school graduate, some college, or college degree), race/ethnicity (Non-Hispanic White, Non-Hispanic Black, Hispanic, or Other), and marital status (married/partnered, divorced/separated, widowed, or never married). These served as baseline characteristics and potential confounders. Age and education are well-established factors related to cognitive reserve and dementia incidence [46,47]. Sex (female) is associated with higher late-life dementia prevalence, which partly reflects longevity and biological factors [46].

Social Engagement: Several measures of social connectedness and activity were included. Social network size was measured as the number of close family/friends with whom the respondent discusses important matters (a proxy for the availability of social support). Community engagement was assessed via self-reported involvement in community groups and the respondent's perception of neighborhood social cohesion/trust. We also included the frequency of leaving the home (ranging from rarely to daily) as an indicator of mobility and social participation in the broader environment. Additionally, social involvement in organized activities such as clubs, classes, or volunteer work was assessed. These variables reflect the hypothesis that an active social life and integration into the community are protective against cognitive decline [12,13,48–51].

Technology Use: This binary composite indicator was created to measure overall digital engagement versus digital exclusion at the population level. While this method enhances model simplicity and scalability, it combines diverse digital behaviors and does not differentiate among specific activities such as communication, information seeking, or financial transactions. Therefore, the significance assigned to this variable reflects general digital engagement rather than any digital activity skill. We coded this as 1 if the person reported using any of these technologies and 0 if they used none. This “Technology use” variable acts as our primary measure of digital engagement in daily life. Prior research suggests that older adults who use information and communication technologies maintain better cognitive function than those who are digitally disconnected [16,52].

Health and Function: We included several self-reported health conditions that are associated with dementia risk. These were combined into composite indicators when appropriate. Vascular conditions were summarized as a binary variable indicating a history of major cardiometabolic diseases, such as heart attack, heart disease, stroke, lung disease, or cancer. We also recorded specific conditions separately: history of stroke, heart disease, hypertension, and diabetes (each coded 0/1). Depressive symptoms were assessed using a brief screening (indicators of frequent sad or hedonic mood); we used a binary variable for clinically relevant depressive symptoms (yes/no) [53]. General health status was assessed using a self-rated overall health scale (5-point scale, ranging from excellent to poor) and a yes/no indicator of mobility difficulties (e.g., concerns about falling) as measures of frailty. We also included smoking status (current smoker or not) as a lifestyle factor [54]. While not all these health variables are primary factors of interest, they help account for physical health differences and comorbidities that could confound the relationships between social and tech engagement and cognition.

To capture both overall comorbidity burden and condition-specific effects, composite indicators, such as `vascular_condition`, were included alongside selected individual diagnoses, such as stroke and hypertension. Although this structure may introduce collinearity in regression models, coefficients were interpreted carefully and not as isolated causal effects. Downstream ML models and SHAP-based explainability analyses, which are less affected by multicollinearity, were used to assess predictor stability.

3.4. Feature Engineering and Data Preparation

All categorical variables, such as education levels and marital status, were one-hot encoded for modeling. Continuous variables (like age) were standardized to z-scores. We handled missing data through a multistep process. First, we removed features that were irrelevant or had too much missing data, as specified in the NHATS documentation. If a participant had more than four key predictors missing, that case was dropped. After initial data cleaning, the remaining missing values in continuous variables were imputed using KNN, while categorical variables were managed through encoding and mode-based strategies. Finally, all features were scaled to the appropriate format (0–1 for binary features, standardized for continuous features, or ordinal as needed) to prepare for model

input. Table 1 presents the definitions, feature names, and measurement types of all study variables.

Table 1. Description and coding of demographic, behavioral, social, technological, and clinical predictor variables used in the analysis.

Variable	Description	Feature Name	Type	Value
Technology Use	Use of computers/tablets, mobile phones, email, internet browsing, online shopping, and online banking	tech_use	Binary	0 = No use, 1 = Regular use
Gender	Self-reported sex	gender	Binary	0 = Female, 1 = Male
Education Level	Highest educational attainment	education	Ordinal	0 = <High school, 1 = High school, 2 = Some college, 3 = College and above
Race/Ethnicity	Self-identified race/ethnicity	ethnicity	Nominal	Categorical
Marital Status	Married/partnered, divorced/separated, widowed, or never married	marital_stat	Nominal	Categorical
Census Division	U.S. Census geographic division	census_division	Nominal	Categorical
Community Engagement	Perceived community trust and cohesion	community	Ordinal	1 = Do not agree, 2 = Neutral, 3 = Agree a lot
Well-being	Emotional and psychological self-assessment	wellbeing	Ordinal	1 = Most of the days, 3 = Rarely
Vascular Condition	Binary indicator of the presence of any vascular comorbidity (heart attack, heart disease, stroke, lung disease, cancer)	vascular_condition	Binary	0 = No vascular condition, 1 = ≥ 1 vascular condition
Overall Health Condition	Self-reported overall health status	overallhealth	Ordinal	1 = Excellent, 5 = Poor
Depressive Symptoms	Symptoms of low mood or loss of interest	depression_cat	Ordinal	1 = None, 5 = Severe
Heart Disease	Self-reported heart disease (e.g., angina, heart failure)	heart_disease	Binary	0 = No, 1 = Yes
Hypertension	Self-reported high blood pressure	hypertension	Binary	0 = No, 1 = Yes
Heart Attack	Self-reported history of myocardial infarction	heart_attack	Binary	0 = No, 1 = Yes
Worry About Falling	Fear of falling affects daily activities	worryfall	Binary	0 = No, 1 = Yes
Fall in Past Year	Self-reported fall within the last year	fallendownyr	Binary	0 = No, 1 = Yes
Vascular Count	Count of reported vascular comorbidities	vascular_count	Count	Integer count (0–N)
Current Smoking	Current smoking status	smokesnow	Binary	0 = No, 1 = Yes
Going Outside	Frequency of leaving home to go outside	go_outside	Binary	0 = Rarely/Never, 1 = Regularly
Number of Social Networks	Number of close contacts for sharing important matters	nsocialnet	Ratio	Count
Diabetes	Self-reported diabetes diagnosis	diabetes	Binary	0 = No, 1 = Yes
Stroke	Self-reported history of stroke	stroke	Binary	0 = No, 1 = Yes
Dementia Onset (Outcome)	Incident dementia during the follow-up period	dementia	Binary	0 = No dementia, 1 = Dementia

Notable predictors include measures of social engagement, such as how often people go outside, the size of their social network, and community cohesion, as well as clinical

indicators like vascular conditions and depressive symptoms, which consistently rank among the most influential features.

3.5. Rationale for Algorithm Selection

The selected algorithms were chosen for their suitability for structured, tabular survey data and their balance between predictive performance and interpretability. Classical and ensemble ML models have proven to perform well on structured population-based health datasets, often surpassing more complex deep learning architectures while maintaining greater transparency and stable calibration. Considering the study's focus on explainability, prevention relevance, and real-world application, these models offer a suitable and reliable foundation for systematic comparison rather than for algorithmic novelty.

4. ML Pipeline and Model Implementation

4.1. ML Algorithms

Five supervised ML algorithms were assessed: Logistic Regression (LR), Support Vector Machine SVM, K-Nearest Neighbors (KNN), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). These models encompass different methodological approaches with varying degrees of complexity, nonlinearity, and interpretability, allowing for systematic comparison using structured survey data.

LR served as a baseline interpretable model, estimating dementia risk as a linear combination of predictors optimized with a log-likelihood loss function and L2 regularization. The SVM used a radial basis function kernel to learn nonlinear decision boundaries by maximizing the margin between classes. The KNN algorithm classified observations based on the majority label among the $k = 5$ nearest neighbors using Euclidean distance.

RF and XGBoost were included as ensemble tree-based methods capable of capturing nonlinear interactions and higher-order feature relationships. Random Forest combined predictions from 100 decision trees built using bootstrap sampling and Gini impurity. XGBoost used gradient-boosted decision trees with a learning rate of 0.1 and a maximum depth of 4, optimizing a regularized objective function to reduce overfitting. Early stopping was employed to prevent unnecessary model complexity.

Hyperparameter values were chosen based on widely reported settings in previous studies, with minimal tuning to ensure comparability across models and reduce the risk of overfitting.

Baseline dementia cases were excluded before analysis. After incorporating engineering and standardization into the training folds, five classifiers (LR, SVM, KNN, RF, and XGBoost) were evaluated using stratified five-fold cross-validation. Incident dementia during follow-up (2012–2022) was predicted. Model performance was assessed with classification metrics and ROC-AUC, and model interpretability was examined using SHAP values, logistic regression odds ratios, and feature importance measures (Figure 2).

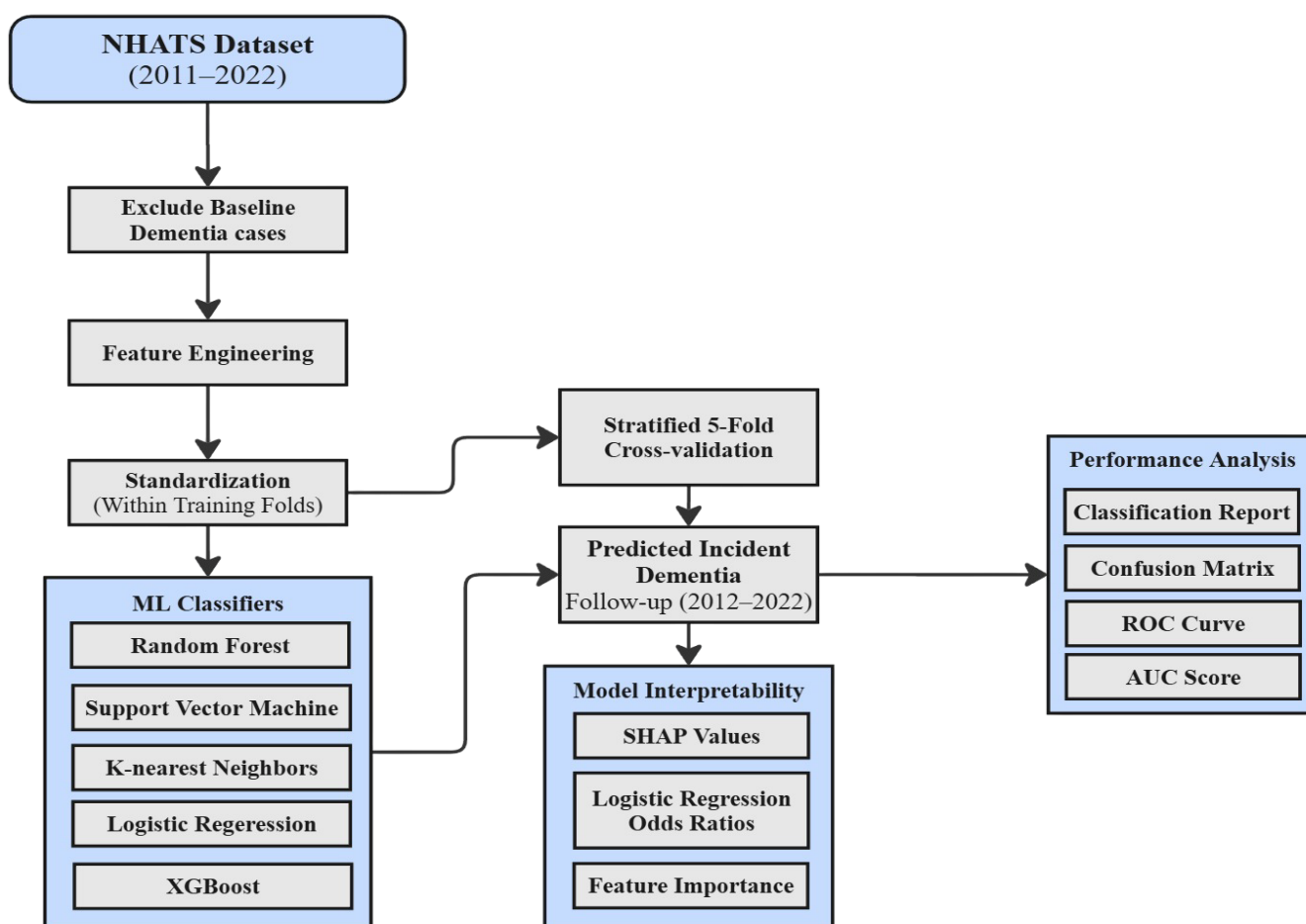


Figure 2. ML pipeline for dementia risk prediction using the National Health and Aging Trends Study (NHATS, 2011–2022).

4.2. Training Procedure and Cross-Validation

All models were trained and evaluated using stratified 5-fold cross-validation to preserve the proportion of dementia and non-dementia cases in each fold. In each iteration, four folds were used for training and one for validation, ensuring every observation served as validation data exactly once. To prevent information leakage, all preprocessing steps, including feature scaling, missing-value imputation, and resampling, were performed only within the training folds and then applied to the corresponding validation fold. Model performance was measured for each fold and averaged across all five folds, providing stable, generalizable estimates of predictive performance.

4.3. Class Imbalance Handling

Dementia cases accounted for about 25% of the study samples, resulting in a somewhat imbalanced outcome distribution. To fix this imbalance, different strategies were used depending on the algorithm. Adaptive Synthetic Sampling (ADASYN) was applied within the training folds for KNN, RF, and XGBoost to generate synthetic minority-class examples and improve sensitivity to dementia cases. For LR and SVM models, class weighting was used to penalize misclassifications of the minority dementia class and to adjust the decision boundary without resampling. No oversampling or weighting was applied to validation data to prevent information leakage and evaluation bias.

4.4. Performance Evaluation Metrics

Model performance was evaluated using both threshold-dependent and threshold-independent metrics to address class imbalance and probabilistic accuracy. Accuracy was included for completeness but was not the sole evaluation metric due to its sensitivity to class imbalance. Imbalance-aware metrics included precision, recall (sensitivity), and the F1 score, which balances false positives and false negatives. Discriminative ability was assessed using the Area Under the Receiver Operating Characteristic Curve (ROC–AUC) and the Area Under the Precision–Recall Curve (PR–AUC). Probabilistic calibration was evaluated with the Brier score, which measures the average squared difference between predicted probabilities and observed outcomes. All metrics were calculated within each cross-validation fold and reported as mean values across folds to ensure a robust and reliable performance assessment.

Accuracy: Accuracy is the proportion of instances correctly classified (both true positives and true negatives) among all observations. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Accuracy provides a general measure of overall performance; however, in the presence of class imbalance, it can be misleading, as high accuracy may be achieved by favoring the majority class.

Precision: Precision evaluates the proportion of predicted positive cases that are truly positive and is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Precision is critical in scenarios where the cost of false positives is high, such as unnecessary clinical follow-ups or diagnostic procedures.

Recall (Sensitivity): Recall reflects the proportion of actual positive cases that the model correctly identifies. It is given by:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Recall is especially critical in medical applications, where false negatives may delay diagnosis or treatment and lead to adverse outcomes.

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure of both. It is computed as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The F1 score is instrumental in imbalanced classification settings, as it accounts for both false positives and false negatives and better reflects performance on the minority class.

Brier Score: The Brier score measures the accuracy of probabilistic predictions and is defined as:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$$

where,

N is the total number of observations,

p_i is the predicted probability of dementia for observation i , and

$y_i \in \{0,1\}$ is the observed outcome (0 = no dementia, 1 = dementia).

Lower Brier scores indicate better-calibrated and more accurate probability estimates.

Area Under the Receiver Operating Characteristic Curve (ROC-AUC): ROC-AUC measures the model's ability to discriminate between classes across all possible classification thresholds. The ROC curve plots the actual positive rate against the false positive rate, and the AUC summarizes the overall discriminative performance. An AUC of 1.0 indicates perfect discrimination, whereas an AUC of 0.5 reflects no discriminative ability.

Area Under the Precision–Recall Curve (PR-AUC): The precision–recall curve plots precision against recall at varying thresholds. PR-AUC is particularly informative in imbalanced datasets, as it focuses on performance for the minority (positive) class and is less influenced by the abundance of true negatives.

Instead of relying solely on accuracy, using these metrics together offers a more detailed and reliable evaluation of model performance. In this study, where the dataset is imbalanced and false negatives have significant clinical consequences, recall and F1 score are especially important. ROC-AUC and PR-AUC provide complementary insights into how well the model differentiates across thresholds, whereas the Brier score assesses the calibration of the predicted probabilities. All reported metrics are the average values from the 5-fold cross-validation process, ensuring stability and reducing variance in performance estimates.

4.5. Model Development and Validation

4.5.1. Data Preprocessing

We performed a comprehensive preprocessing pipeline before training the model. Features with excessive missingness (>30%) or low relevance were initially removed from the dataset. Participants missing many key predictors were excluded to ensure data quality. For the remaining missing entries, we used appropriate imputation methods to maintain local data relationships. All categorical variables were then encoded as binary indicators for dichotomous features and as one-hot encodings for multi-class features, while ordinal variables were retained as ordered numeric codes. Continuous variables were standardized to z-scores within each training fold and then applied to the corresponding validation fold to prevent information leakage. We did not use dimensionality-reduction techniques because our goal was to keep the interpretability of all predictors. Class imbalance was managed during model training with algorithm-specific techniques. ADASYN was used within training folds for KNN, RF, and XGBoost to create synthetic minority-class samples, while LR and SVM used class weighting to increase the penalty for misclassifying dementia cases. Oversampling and weighting were only applied to the training subsets within each cross-validation fold to avoid data leakage into the validation set. This approach resulted in balanced training data while maintaining unbiased model evaluation.

4.5.2. Models and Training Procedure

We evaluated five supervised learning algorithms commonly used for classification: LR, SVM, KNN, RF, and XGB. Models were configured using literature-supported default hyperparameters with minimal tuning to maintain comparability and reduce overfitting risk. The SVM utilized a radial basis function (RBF) kernel with $C = 1.0$; KNN used $k = 5$ neighbors. The RF ensemble consisted of 100 decision trees, and XGB was trained with a maximum tree depth of 4 and a learning rate of 0.1. For XGB, an early-stopping criterion of 50 consecutive rounds without improvement was employed to help prevent overfitting.

All models were trained and validated using the same feature set and data splits to ensure a fair comparison. Model development involved stratified 5-fold cross-validation, in which the dataset was divided into 5 folds while preserving class proportions. In each iteration, four folds were used for training and one for validation; this process was repeated so that each observation served exactly once as validation data. Oversampling was limited to adaptive methods (ADASYN) within training folds only; no global or validation-level resampling was conducted. The final model performance was reported as the average across the five validation folds.

4.5.3. Performance Evaluation

We evaluated model performance using several complementary metrics: accuracy, F1-score, ROC AUC, and Brier score [9]. Accuracy measures overall correctness, while the F1-score balances sensitivity and precision, making it especially informative when the positive and negative classes are not equally distributed. The ROC AUC (Area Under the Receiver Operating Characteristic curve) summarizes the model's ability to discriminate across different thresholds, and the Brier score evaluates the mean squared error of probabilistic predictions, reflecting calibration accuracy. Each metric was calculated for each fold, and we report the mean $\pm$ standard deviation across the 5 folds as the overall performance. To compare models statistically, we analyzed the distribution of fold-level metrics. The Shapiro–Wilk test was used to assess the normality of performance metrics across folds. Based on the normality results, suitable statistical tests were selected. We also validated the findings with visual diagnostics: ROC curves were plotted for each model to evaluate classification trade-offs, and confusion matrices were examined for patterns in false positives and negatives. Additionally, we created boxplots and histograms of fold-wise metrics to illustrate variability and support the results. These visual and statistical checks confirmed the robustness of our cross-validation outcomes.

4.5.4. Reproducibility and Interpretability

To ensure reproducibility, all modeling steps were carried out using a consistent pipeline across models, from preprocessing to cross-validation, with fixed random seeds for data splitting and model initialization where applicable. Each algorithm was evaluated on the same data splits, using identical training data, to enhance the validity of performance comparisons. Additionally, interpretability was a key aspect of our model development. We employed SHAP (Shapley Additive exPlanations) analysis as a post hoc interpretability method for the tree-based models (RF and XGB). SHAP assigns an essential score to each feature for individual predictions, enabling interpretation of model outputs in terms of feature contributions. Using SHAP, we confirmed that the top features influencing dementia predictions in the ensembles align with domain knowledge; for example, technology use and social engagement indicators consistently appeared as high-impact predictors in both RF and XGB models.

5. Hypotheses

To operationalize the study's four research questions, we formulate the following testable hypotheses:

5.1. Hypothesis 1 (H1): Behavioral and Technological Predictors

This hypothesis, addressing RQ1, tests whether modifiable behavioral and technological factors are significantly associated with dementia risk. A multiple logistic regression model was used to evaluate the predictive value of features such as technology use, social network size, and frequency of going outside.

H₀: Behavioral and technological factors such as digital engagement and social participation are not associated with dementia risk (all coefficients = 0).

H₁: At least one behavioral or technological factor is significantly associated with dementia risk (at least one coefficient $\neq 0$).

5.2. Hypothesis 2 (H2): Model Performance Relative to Random Baseline

This hypothesis addresses Research Question 2 (RQ2), which investigates whether ML models provide probabilistic predictions that exceed those of a non-informative reference. Predictive performance is evaluated using the Brier score (μ), where lower scores signify better probabilistic calibration.

For each model, the null hypothesis assumes that performance equals a prevalence-based non-informative baseline Brier score derived from the observed dementia prevalence in the study sample. Let μ_0 denote the prevalence-based baseline Brier score. Let μ denote the prevalence-based baseline Brier score, defined as $\mu_0 = p(1 - p)$, where p is the observed dementia prevalence in the study sample.

$$\mu_0 = p(1 - p),$$

where p is the observed dementia prevalence. This baseline corresponds to a non-informative model that predicts a constant probability equal to the empirical event rate for all observations.

H2a–H2e: One-Sample Tests for Individual Models

$$H0a \text{ (RF): } \mu_{\text{RF}} \geq \mu_0 \quad H1a: \mu_{\text{RF}} < \mu_0$$

$$H0b \text{ (XGB): } \mu_{\text{XGB}} \geq \mu_0 \quad H1b: \mu_{\text{XGB}} < \mu_0$$

$$H0c \text{ (LR): } \mu_{\text{LR}} \geq \mu_0 \quad H1c: \mu_{\text{LR}} < \mu_0$$

$$H0d \text{ (SVM): } \mu_{\text{SVM}} \geq \mu_0 \quad H1d: \mu_{\text{SVM}} < \mu_0$$

$$H0e \text{ (KNN): } \mu_{\text{KNN}} \geq \mu_0 \quad H1e: \mu_{\text{KNN}} < \mu_0$$

Rejecting the null hypothesis shows that the model's probabilistic predictions perform better than a non-informative baseline suitable for an uneven outcome distribution.

5.3. Hypothesis 3 (H3): Feature Importance Patterns Across Models

This hypothesis addresses Research Question 3 (RQ3), which examines whether the most influential predictors of dementia risk differ across algorithms and if certain behavioral and technological features consistently remain essential.

H3a: Variation in Feature Importance by Model

- **H₀:** Feature importance values do not significantly differ across models.
- **H₁:** At least one model assigns significantly different importance rankings to predictors.

H3b: Consistency of Key Predictors

- **H₀:** Behavioral and technological features such as technology use, how often people go outside, and the size of social networks are not consistently ranked as the top features.

- **H₁:** These modifiable predictors consistently appear among the top-ranked features across all models.

Together, these hypotheses examine both the variability of feature importance across algorithms (H3a) and the consistency of key dementia-related predictors across models (H3b), directly addressing RQ3.

5.4. Hypothesis 4 (H4): Cross-Model Calibration Differences

This hypothesis assesses whether probabilistic calibration varies among models by using fold-level Brier scores and nonparametric repeated-measures tests. It addresses Research Question 4 (RQ4), which examines whether explainable machine learning models can produce reliable, interpretable predictions that support innovative, sustainable healthcare strategies.

- **H₀:** All models provide equally calibrated predictions (no difference in Brier scores).
- **H₁:** At least one model provides better calibration (lower Brier score).

Interpretability was further supported by SHAP analysis, which revealed that features such as technology use and social engagement consistently influenced predictions across models.

6. Results and Analysis

6.1. Model Performance Comparison

This section summarizes the experimental results across the evaluated algorithms. As shown in Table 2, XGBoost demonstrated the strongest overall performance across both discrimination and metrics robust to uneven class distributions. Although XGB and RF achieved the highest classification accuracy (0.881 and 0.878, respectively), accuracy alone did not fully reflect the effectiveness of these models under disproportionate outcome representation. XGB achieved the highest ROC–AUC (0.823) and a competitive F1-score (0.435), indicating a stronger ability to distinguish dementia cases while maintaining balanced precision–recall performance.

Table 2. Cross-validated performance metrics of ML models (mean $\pm$ standard deviation).

Model	Accuracy (Mean $\pm$ SD)	Mean Exec. Time [ms] (Mean $\pm$ SD)	F1 Score (Mean $\pm$ SD)	Brier Score (Mean $\pm$ SD)	ROC-AUC (Mean $\pm$ SD)
KNN	0.689 $\pm$ 0.003	56.300 $\pm$ 39.161	0.377 $\pm$ 0.004	0.228 $\pm$ 0.001	0.717 $\pm$ 0.004
LR	0.746 $\pm$ 0.006	640.256 $\pm$ 95.533	0.448 $\pm$ 0.011	0.174 $\pm$ 0.002	0.813 $\pm$ 0.006
XGB	0.881 $\pm$ 0.004	1291.280 $\pm$ 295.747	0.435 $\pm$ 0.013	0.094 $\pm$ 0.002	0.823 $\pm$ 0.005
RF	0.878 $\pm$ 0.004	18,522.245 $\pm$ 734.812	0.432 $\pm$ 0.013	0.096 $\pm$ 0.001	0.818 $\pm$ 0.004
SVM	0.772 $\pm$ 0.007	568,935.469 $\pm$ 28281.250	0.469 $\pm$ 0.010	0.098 $\pm$ 0.001	0.817 $\pm$ 0.003

LR and SVM demonstrated moderate accuracy (0.746 and 0.772, respectively) but comparatively higher F1-scores (0.448–0.469), suggesting improved sensitivity to dementia cases relative to KNN. In contrast, KNN, despite its computational efficiency, showed the weakest performance on imbalance-sensitive metrics, with the lowest F1-score (0.377) and ROC–AUC (0.717), reflecting a tendency to favor the majority non-dementia class. Overall, these findings highlight the robustness of ensemble-based approaches, particularly XGB, in achieving well-balanced predictive performance across multiple evaluation metrics. Figure 3 demonstrates the stability of accuracy, F1 score, Brier score, and ROC AUC across the five stratified cross-validation folds.

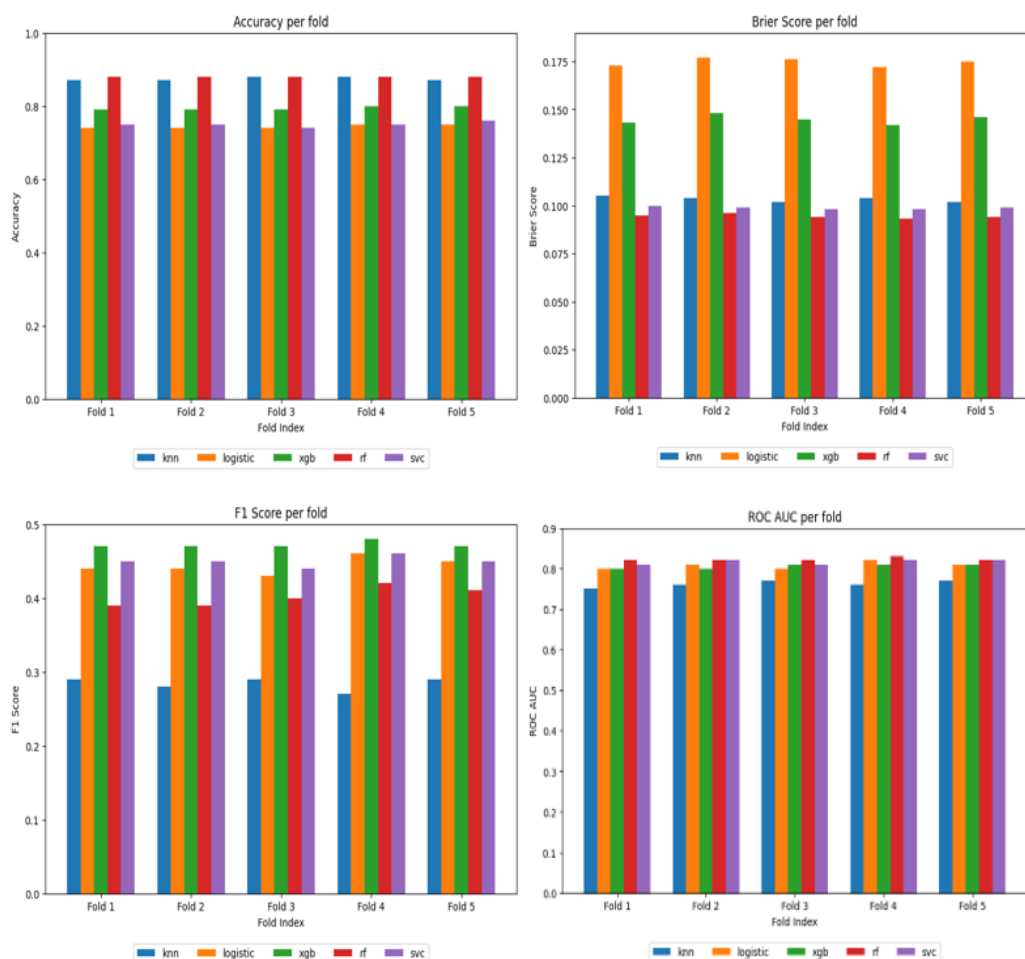


Figure 3. Cross-Validation Performance of ML Models Across 5 Folds.

Although accuracy is reported for completeness, the model evaluation focused mainly on metrics less affected by uneven class proportions, such as the F1 score, ROC AUC, and Brier score, since these offer a more reliable measure of predictive performance when the outcome distribution is uneven.

Table 3 presents the results of the Shapiro–Wilk test for assessing the normality of feature-importance distributions across five machine-learning algorithms. Several features, particularly in the XGB and RF models (hypertension and vascular condition), display significant deviations from normality ($p < 0.05$). Because of these violations, parametric comparisons were considered unsuitable, so nonparametric statistical tests were used for cross-model comparisons of feature importance.

Table 3. Shapiro–Wilk test p -values for evaluating the normality of feature-importance scores across five machine-learning algorithms.

Feature	KNN	LR	XGBoost	RF	SVC
gender	0.6027	0.0741	0.8271	0.1219	0.0741
hypertension	0.7098	0.1405	0.0375	0.0186	0.1405
go_outside	0.9674	0.1695	0.3624	0.0989	0.1695
census_division	0.2366	0.1890	0.8062	0.6895	0.1890
nsocialnet	0.8691	0.1925	0.8515	0.4520	0.1925
social_participation	0.4457	0.1939	0.9125	0.4627	0.1939
worryfall	0.6260	0.2697	0.0959	0.3814	0.2697
community_score	0.3098	0.3863	0.4537	0.5803	0.3863
ethnicity	0.9789	0.4931	0.9682	0.6481	0.4931

community	0.5073	0.5394	1.0000	0.3798	0.5394
vascular_condition	0.0021	0.6532	1.0000	0.9653	0.6532
overallhealth	0.8100	0.6251	0.4695	0.1164	0.6251
heart_disease	0.8706	0.6633	0.5750	0.6116	0.6633
depression	0.3352	0.6321	0.1776	0.9220	0.6321

Note: Bold values indicate violations of normality assumptions.

The overall cross-validated performance metrics averaged across folds are summarized in Figure 4. The average cross-validated performance metrics (accuracy, F1 score, Brier score, and ROC–AUC) of five ML algorithms. The values reflect the averages reported in Table 2; error bars are left out for clarity. Likewise, XGBoost exhibits higher ROC–AUC values than the other classifiers, indicating stronger discrimination performance (Figure 4). Beyond overall performance and stability, model discrimination ability was further examined using receiver operating characteristic (ROC) curves. Figure 5 presents the receiver operating characteristic (ROC) curves across a stratified five-fold cross-validation for five ML models: KNN, LR, SVM, XGBoost, and RF. Each subplot displays fold-specific ROC curves and their corresponding AUC values.

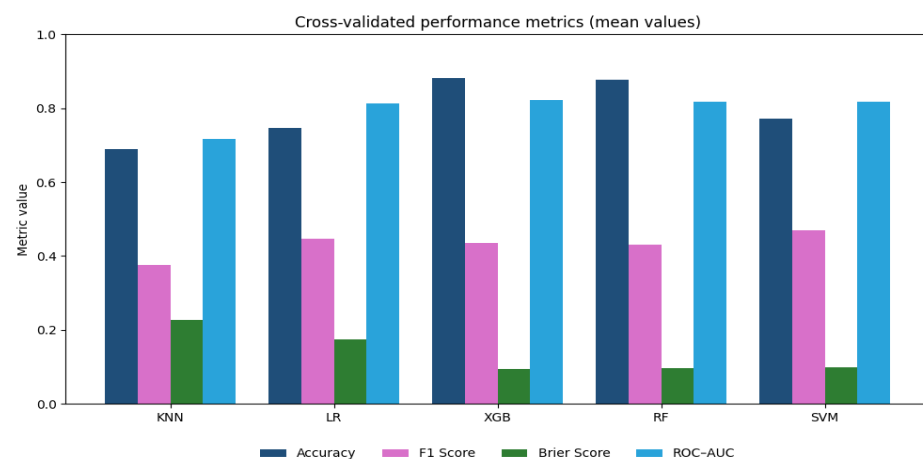


Figure 4. Bar chart of performance metrics of ML algorithms.

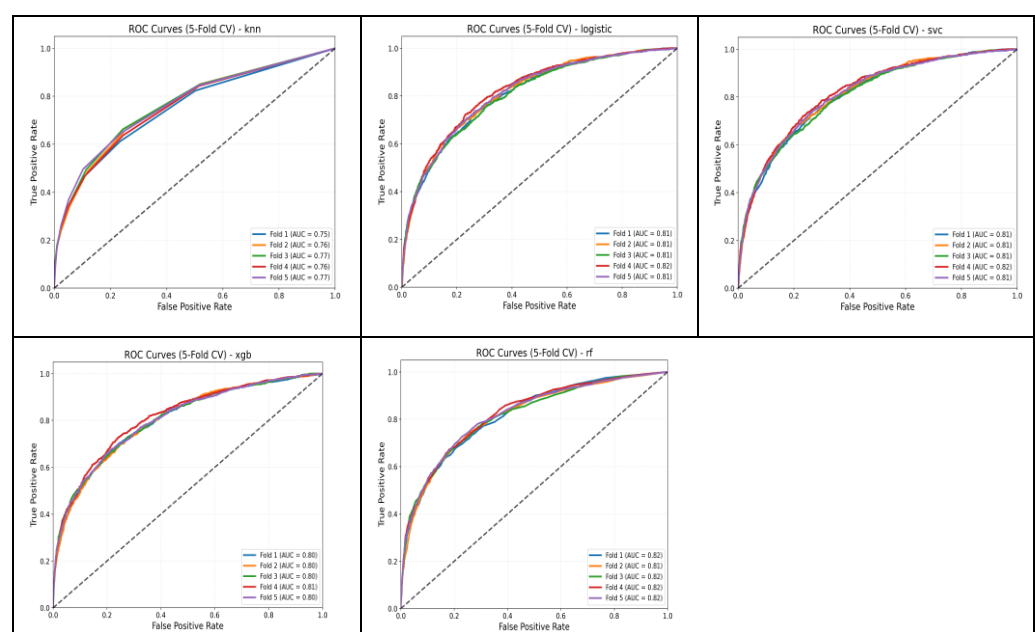


Figure 5. ROC curves illustrating model discrimination across five stratified cross-validation folds for five ML algorithms: KNN, LR, SVM, XGBoost, and RF. Each panel shows fold-level ROC curves

and corresponding AUC values. The diagonal dotted line represents the reference line corresponding to random classification ($AUC = 0.5$).

Across folds, XGBoost consistently achieved the highest ROC–AUC values (0.82–0.85 per fold), indicating stronger discrimination than LR (0.80–0.81) and KNN (0.75–0.77). The ROC curves for XGBoost are closer to the top-left corner, reflecting a more favorable trade-off between true-positive and false-positive rates. RF showed comparable discrimination, with ROC–AUC values slightly below those of XGBoost. In contrast, KNN showed weaker discrimination, as reflected by lower ROC–AUC scores and less robustness when class proportions were unequal. Logistic Regression showed moderate discrimination but the weakest calibration, as indicated by its higher Brier score. SVM achieved relatively good discrimination but did not match the ensemble models, despite exhibiting better calibration than LR.

The ROC analyses support the tabulated results, indicating that ensemble methods, especially XGBoost, provide the strongest discrimination performance, followed closely by RF, while linear and distance-based classifiers show more limited performance in this dementia prediction task with uneven outcome distribution. To thoroughly compare model performance, we conducted statistical tests on the cross-validation results. Since performance metrics were obtained from the same cross-validation folds, the observations were not independent. Therefore, calibration differences were evaluated using the Friedman test (see Section 6.2)

6.2. Statistical Testing

Because all algorithms were evaluated using the same stratified five-fold cross-validation splits, fold-level performance metrics reflect dependent repeated measures rather than independent observations. Therefore, parametric tests such as one-way ANOVA, which assume independence, were not used. Model performance comparisons are thus described with cross-validated summaries of discrimination and calibration metrics (ROC–AUC and Brier score), supported by plots showing fold-level variability. When formal hypothesis testing was needed, nonparametric methods suitable for dependent samples were employed, and the results were interpreted as evidence of consistent cross-validation performance patterns rather than population-level conclusions.

6.2.1. Model Interpretability

To understand the drivers of the models' predictions, we analyzed feature importance using SHAP (Shapley Additive Explanations) and examined the logistic regression coefficients. Figures 6 and 7 support Hypotheses 1 and 3b, key behavioral and social factors, such as technology use, social network size, and frequency of going outside, consistently rank among the most influential predictors across models. These patterns underscore their importance in predicting dementia risk. The SHAP summary plots in Figures 6 and 7 highlight the contribution of each feature to individual dementia risk predictions. Technology use, frequency of outdoor activity, and social network size consistently rank among the top three predictors across models.

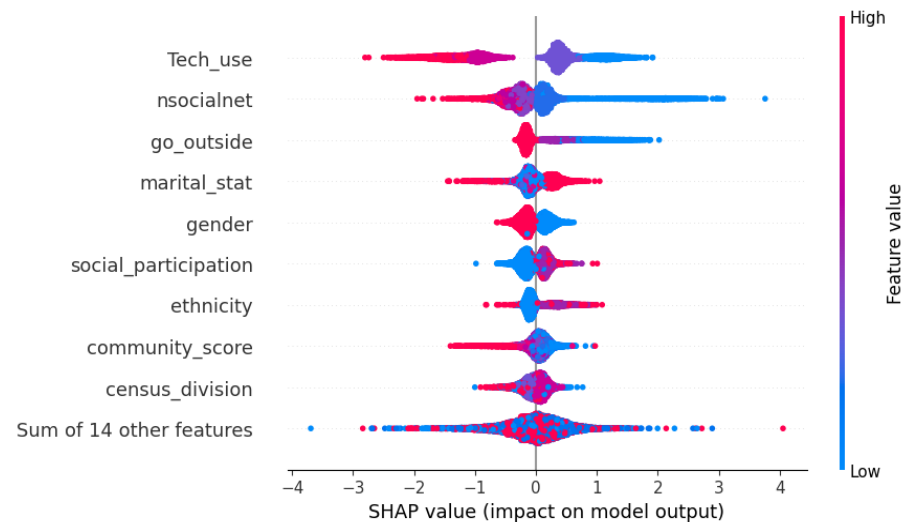


Figure 6. SHAP summary plot of top features on dementia prediction.

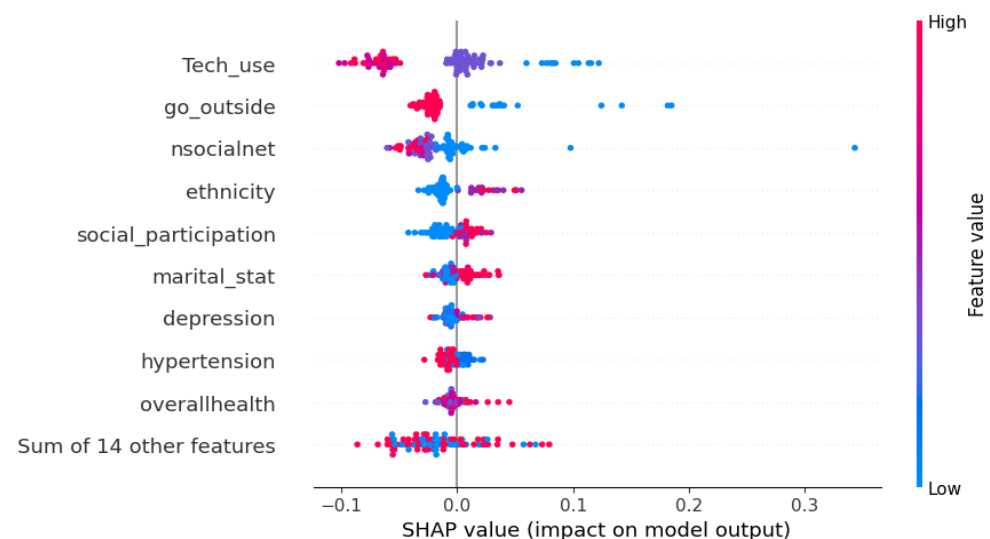


Figure 7. SHAP summary plot of feature contributions to dementia prediction.

The SHAP violin plots indicate that technology use, frequency of going outside, and social network size exhibit the widest distributions, suggesting a strong predictive influence. Higher values of these features (red points) are associated with negative SHAP values, indicating lower predicted dementia risk, whereas lower values (blue points) are associated with higher risk. These patterns closely align with the logistic regression results: technology use (OR = 0.47), outdoor activity (OR = 0.66), and social network size (OR = 0.63) demonstrate significant protective effects ($p < 0.001$), with confidence intervals reported on the odds ratio scale. In contrast, marital status (OR = 1.13), social participation (OR = 1.12), and non-White ethnicity (OR = 1.19) are associated with higher predicted risk, likely reflecting contextual or unmeasured confounding factors. Figure 5 shows the average absolute SHAP values for key features across five models: KNN, LR, XGB, RF, and SVM. Technology use consistently emerges as the most influential predictor across models, with a normalized SHAP value of 1.000. Since technology use was modeled as a binary composite indicator, this finding highlights the importance of overall digital engagement rather than specific digital activities, supporting H3.

Figure 8 shows a heatmap of average absolute SHAP values for each feature. Darker colors indicate higher importance. The heatmap demonstrates that a core group of features, especially technology use, frequency of going outside, and social network size,

consistently rank among the top predictors across all models. Technology use is the most influential feature in four of the five models (except KNN, where it ranks second after frequency of going outside). In contrast, both the frequency of going outside and the size of the social network appear among the top three for every algorithm. This agreement across models is notable, given the diversity of model types, ranging from LR to nonlinear ensemble methods such as XGB and RF.

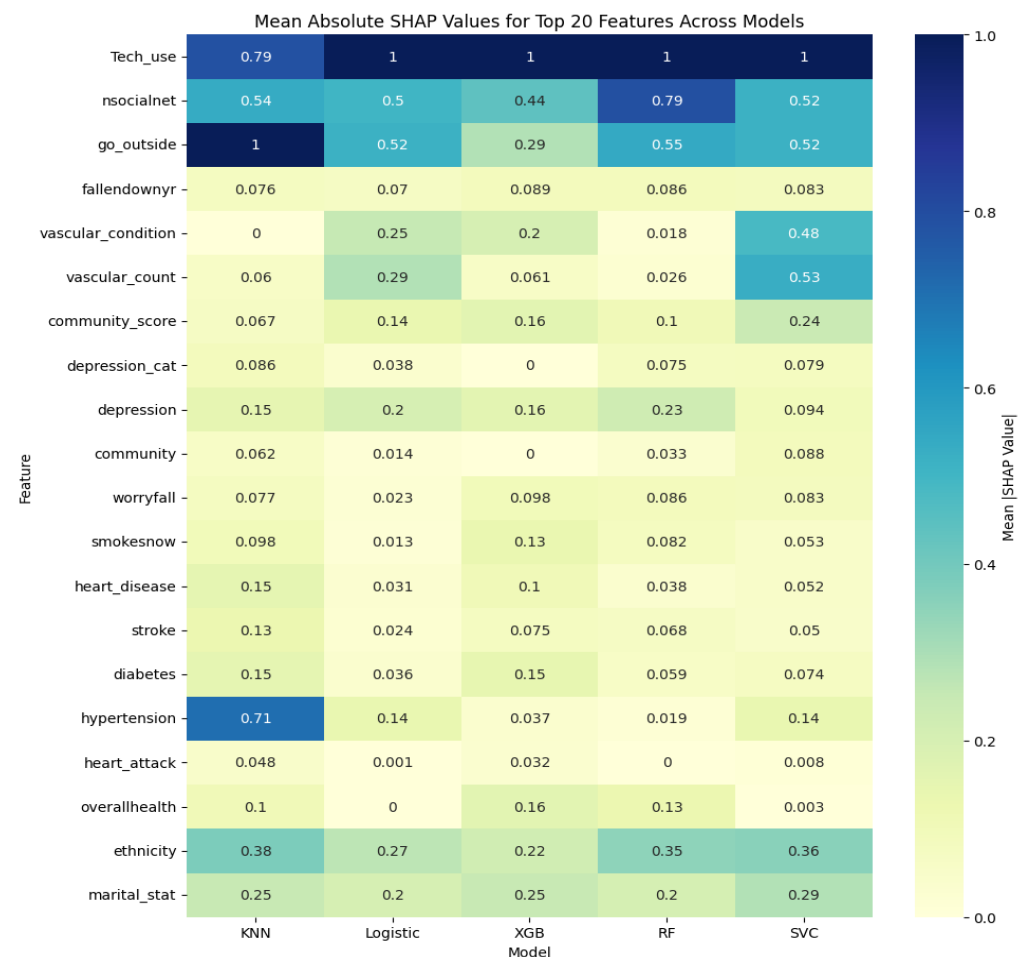


Figure 8. SHAP features importance heatmap across models.

6.2.2. Hypothesis 1 (H1): Behavioral/Social Predictors and Dementia Risk

Table 4 summarizes the results of the multiple logistic regression analyzing the relationships between behavioral, social, and demographic factors and the onset of dementia. The overall model was statistically significant ($\chi^2(20) = 260.58$, $p < 0.001$) and explained about 9% of the variance in dementia outcomes (pseudo- $R^2 = 0.09$). Out of the 20 predictors included, 12 were statistically significant ($p < 0.05$). Regular technology use, more frequent outdoor activities, and larger social networks were associated with lower odds of dementia ($OR < 1$), indicating protective effects. Conversely, marital status, racial/ethnic minority status, and participation in structured social activities were linked to higher odds of dementia ($OR > 1$). These findings support H1 by showing that several behavioral and social engagement factors are significantly related to dementia risk among older adults.

Table 4. Logistic Regression Results: Predictors of Dementia Risk.

Feature	Odds Ratio (OR)	95% CI (OR)	<i>p</i> -Value
Tech_use	0.47	(0.45, 0.50)	<0.001
go_outside	0.66	(0.63, 0.68)	<0.001
nsocialnet	0.63	(0.60, 0.66)	<0.001
Ethnicity	1.19	(1.15, 1.23)	<0.001
marital_stat	1.13	(1.08, 1.18)	<0.001
social_participation	1.12	(1.07, 1.16)	<0.001
Gender	0.89	(0.85, 0.93)	<0.001
fallendownyr	0.92	(0.89, 0.95)	<0.001
Stroke	0.95	(0.93, 0.98)	0.001
hypertension	0.93	(0.89, 0.97)	0.002
census_division	1.05	(1.01, 1.10)	0.007
Diabetes	0.95	(0.92, 0.99)	0.013
overallhealth	0.96	(0.92, 1.00)	0.057
depression	1.10	(0.98, 1.22)	0.113
heart_attack	0.98	(0.95, 1.01)	0.177
Worryfall	1.03	(0.99, 1.06)	0.186
heart_disease	1.02	(0.98, 1.06)	0.337

Notes: Odds ratios (OR) and 95% confidence intervals are reported on the same scale for interpretability. OR < 1 indicates a protective association (lower odds of dementia), whereas OR > 1 indicates an increased association (higher odds of dementia). Confidence intervals that include 1 indicate non-significant associations at $\alpha = 0.05$. Statistically significant results ($p < 0.05$) are shown in bold.

6.2.3. Hypothesis 2 (H2): Model Performance vs. Chance

To assess whether each model's probabilistic predictions surpassed a meaningful reference, we compared the mean Brier score of each algorithm against a prevalence-based non-informative baseline derived from the observed dementia rate in the study sample. Because the outcome is imbalanced (approximately 25% dementia prevalence), this baseline corresponds to predicting a constant probability equal to the empirical event rate for all observations, rather than assuming a balanced outcome. One-sample one-tailed tests were performed after evaluating distributional assumptions with the Shapiro–Wilk test; when normality was violated, nonparametric tests were used. As shown in Table 5, all five ML models (LR, SVM, KNN, RF, and XGB) achieved mean Brier scores significantly lower than the prevalence-based baseline ($p < 0.001$), thereby rejecting the null hypothesis for each algorithm. These results demonstrate that all models produced probabilistic predictions that outperformed a non-informative reference by a considerable margin.

Table 5. One-sample tests comparing model Brier scores against a prevalence-based baseline derived from the observed dementia rate.

Algorithm	<i>p</i> -Value ($\mu \neq \mu_0$)	<i>p</i> -Value ($\mu < \mu_0$)	<i>p</i> -Value ($\mu > \mu_0$)
KNN	1.037482×10^{-9}	5.187412×10^{-10}	1.0
LR	1.503313×10^{-7}	7.516564×10^{-8}	1.0
XGB	5.617123×10^{-8}	2.808561×10^{-8}	1.0
RF	1.304699×10^{-9}	6.523494×10^{-10}	1.0
SVM	2.354709×10^{-10}	1.177355×10^{-10}	1.0

Note: μ_0 denotes the prevalence-based non-informative baseline Brier score, defined as $\mu_0 = p(1 - p)$, where p is the observed dementia prevalence in the study sample.

6.2.4. Hypothesis 3 (H3): Variation and Consistency in Feature Importance

To examine how different models rank predictive features, we used the Kruskal–Wallis test on SHAP value distributions across five algorithms (KNN, LR, XGB, RF, and SVM). Although SHAP is typically used with tree-based models, SHAP values for logistic regression were computed using a linear explainer, enabling consistent comparisons across models. As shown in Table 6, the test revealed statistically significant differences in feature importance distributions among the algorithms ($p < 0.05$), indicating variation in how models prioritize individual predictors. Since several feature importance distributions violated normality assumptions, the Kruskal–Wallis test served as a suitable non-parametric alternative for comparing feature rankings across models.

Table 6. Kruskal–Wallis Test on Top 20 SHAP Features Across Models (H3).

Feature	<i>p</i> -Value	χ^2	KNN	LR	XGB	RF	SVM
Tech_use	<0.001 **	17.86	55.0	90.0	40.0	15.0	65.0
nsocialnet	<0.001 **	17.86	58.0	90.0	15.0	65.0	65.0
go_outside	<0.001 **	17.86	52.0	90.0	40.0	15.0	65.0
fallendownyr	<0.001 **	16.21	63.0	81.0	15.0	40.0	74.0
vascular_condition	<0.001 **	16.46	72.0	74.0	15.0	40.0	84.0
vascular_count	<0.001 **	16.14	60.0	75.0	40.0	15.0	65.0
community_score	<0.001 **	16.90	54.0	69.0	15.0	70.0	60.0
depression_cat	0.005 **	12.98	64.0	65.0	40.0	15.0	76.0
depression	0.023 *	9.570	57.0	71.0	19.0	67.0	53.0
community	0.002 **	14.95	61.0	59.0	15.0	60.0	70.0
worryfall	0.007 **	12.23	59.0	62.0	45.0	61.0	70.0
smokesnow	0.002 **	14.73	53.0	46.0	59.0	40.0	55.0
heart_disease	<0.001 **	16.55	70.0	71.0	15.0	40.0	84.0
stroke	<0.001 **	17.10	74.0	68.0	40.0	15.0	87.0
diabetes	<0.001 **	16.42	69.0	72.0	15.0	40.0	83.0
hypertension	<0.001 **	15.96	71.0	70.0	18.0	37.0	85.0
heart_attack	0.003 **	14.18	66.0	43.0	76.0	16.0	75.0
overallhealth	0.007 **	12.13	62.0	38.0	52.0	90.0	30.0
ethnicity	<0.001 **	16.90	68.0	69.0	40.0	15.0	86.0
marital_stat	<0.001 **	17.58	56.0	66.0	15.0	45.0	69.0

Note: χ^2 values are reported from the Kruskal–Wallis test. * $p < 0.05$; ** $p < 0.01$. Rank values reflect SHAP feature ranking across models.

Despite statistically significant differences in feature rankings across models, a core set of predictors, particularly technology use, social network size, and frequency of outdoor activity, consistently appears among the top features in at least some algorithms. This pattern confirms H3a by illustrating variability in feature importance across modeling methods and supports H3b by demonstrating that key behavioral and social predictors remain substantively relevant despite methodological differences. Collectively, these findings improve the interpretability and practical relevance of the proposed multi-model framework for dementia risk prediction.

6.2.5. Hypothesis 4 (H4): Differences in Predictive Performance

To evaluate H4, we examined whether probabilistic calibration varied across models using fold-level Brier scores obtained from stratified cross-validation. Since performance estimates were derived from the same folds, dependence among observations was addressed with the Friedman test for related samples. The test showed a statistically significant overall difference in calibration performance among the five models ($\chi^2(4) = 20.00$, p

< 0.001), indicating systematic differences in Brier score rankings across cross-validation folds. Descriptively, ensemble models (XGBoost and RF) and SVM consistently achieved lower Brier scores than LR and KNN across all folds, indicating better calibration. Although post hoc Wilcoxon signed-rank tests did not remain statistically significant after Holm correction, likely due to the limited number of folds ($n = 5$), the significant global test, consistent ranking pattern, and large effect size together support a stable and robust ordering of model calibration performance across cross-validation folds. These findings support H4.

7. Discussion

7.1. Behavioral, Social, and Technology-Related Factors Associated with Dementia

The logistic regression analysis indicates that behavioral and social engagement factors are significantly associated with dementia status, although the model's overall explanatory power remains modest (pseudo- $R^2 = 0.09$). Regular technology use, more frequent outdoor activities, and larger social networks were indicating that engagement-related behaviors are crucial for preserving cognitive health. Several predictors did not reach statistical significance, which may partly result from collinearity among correlated lifestyle and health variables, as discussed in the Methods, as well as the complex, multifactorial nature of dementia [38]. These findings align with evidence from large-scale meta-analyses and cohort studies. Digital engagement via computers, smartphones, and the internet has been associated with a significantly lower risk of cognitive decline, supporting the technological reserve hypothesis, which posits that cognitively stimulating activities enhance resilience to neurodegeneration [55]. Similarly, ongoing social engagement and robust social networks have been shown to reduce dementia risk by strengthening cognitive reserve and buffering psychosocial stress [56].

Together, the regression results and prior research suggest that modifiable behavioral and social factors influence dementia risk in later life. Although individual effect sizes are moderate, their combined public health impact is significant, emphasizing the importance of policies and interventions that foster digital inclusion, social participation, and community engagement among older adults.

7.2. Comparative Performance of ML Models

These findings demonstrate how AI techniques, particularly ensemble learning methods, can capture complex nonlinear behavioral patterns that traditional statistical models may overlook in population-level dementia risk assessment. A comparison of five ML algorithms revealed meaningful differences in their predictive performance using behavioral and social indicators. XGBoost achieved the strongest overall performance, with the lowest Brier score and highest ROC-AUC, while SVM attained the highest F1 score, indicating improved sensitivity to dementia cases. These differences were evaluated using cross-validated calibration and discrimination metrics, with nonparametric tests applied when appropriate for dependent samples. The strong performance of XGB and RF aligns with prior research indicating that tree-based ensemble methods effectively capture nonlinear relationships and higher-order interactions commonly observed in behavioral health data [57]. In contrast, LR and SVM demonstrate weaker calibration and lower sensitivity, reflecting known limitations of linear and margin-based classifiers when modeling complex, non-linear risk patterns [57]. Although KNN achieves relatively high accuracy, its low F1 score indicates limited effectiveness in detecting dementia cases, underscoring the need to evaluate multiple performance metrics in imbalanced classification settings. Overall, these findings suggest that ensemble models, particularly XGB, are

better suited for dementia risk prediction in community-dwelling populations, offering a good balance among accuracy, calibration, and interpretability.

7.3. Key Predictors of Cognitive Decline Across Models

Feature-importance analyses across all five machine-learning models reveal both variability and convergence in the relevance of predictors. While the magnitude of feature contributions varies across algorithms, several variables consistently emerge as influential across models. Technology use, frequency of outdoor activity, and social network size are among the top predictors in nearly all models, as evidenced by SHAP summary plots and feature-importance heatmaps.

Demographic factors such as age and education also demonstrate stable importance across models, consistent with epidemiological evidence identifying age as the primary non-modifiable dementia risk factor and education as a marker of cognitive reserve [46,47]. Rather than reporting associations from a single modeling approach, this study demonstrates that behavioral and social predictors consistently retain their influence across different ML algorithms [58]. Overall, the consistency of key behavioral and social variables across different modeling approaches shows that these variables reliably indicate cognitive decline. Using explainable ML techniques further allows transparent interpretation of these relationships, supporting early risk detection and targeted interventions to promote healthy cognitive development and aging [59–62].

7.4. Implications for Interpretable and Actionable Dementia Risk Assessment

This study demonstrates how AI, in combination with interpretable ML techniques, can integrate behavioral, social, and technological factors into scalable, transparent tools for assessing dementia risk. The consistent identification of modifiable predictors, especially technology use, outdoor activity, and social network size across both LR and SHAP-based analysis, highlights the ability of explainable AI to go beyond prediction and offer actionable insights for dementia risk assessment [63]. By offering clear explanations of how specific behaviors affect individual risk estimates, these models can guide targeted interventions and personalized risk communication.

From a public health and clinical perspective, using interpretable models such as XGB and RF enables scalable risk stratification while preserving transparency and trust in algorithms' decision-making [64]. These methods are beneficial for early screening and prevention, in which non-clinical, survey-based data often serve as the primary source for large-scale population studies. Therefore, combining explainable ML with population-level aging data enables the development of decision-support tools that are accurate, understandable, and actionable, aligning with current research priorities in digital health and cognitively relevant predictive modeling [65,66].

7.5. Limitations and Future Research

This study has several limitations that should be kept in mind when interpreting the results. First, the analysis mainly relies on self-reported survey data from NHATS, which may be affected by recall bias, reporting inaccuracies, and social desirability effects, especially for behavioral and technology-use variables. Although NHATS employs validated instruments and rigorous data-collection protocols, measurement error cannot be fully eliminated. Similar limitations are common in large population-based aging surveys that depend on self-reported measures of social engagement and digital behavior [67,68]. Second, although dementia onset was tracked over time, several predictors were measured only at a single point or at infrequent intervals. This limits the ability to determine the order of changes in behavioral factors and dementia risk, so causal inferences should be avoided. The reported associations should be viewed as predictive rather than causal.

Third, although interpretable ensemble models like XGBoost and RF showed strong predictive performance, results might vary with different feature engineering strategies or hyperparameter settings not tested in this study [63,64,69,70]. In addition, class imbalance and potential underdiagnosis of dementia in survey-based data may affect model generalizability, particularly for underrepresented or vulnerable subpopulations [71]. Fourth, the LR part of the analysis may be affected by collinearity among correlated health indicators, as explained in the Methods. To address this issue, regression coefficients were carefully interpreted, and the main conclusions were based on cross-validated ML performance and SHAP-based explainability, which are less affected by multicollinearity.

Finally, although the models showed stable performance during stratified cross-validation, external validation with independent cohorts was not performed. As a result, the applicability of the findings beyond the NHATS sampling frame, including non-U.S. populations and institutionalized older adults, still needs to be confirmed.

Future research could build on this work by exploring how dementia risk prediction models perform as behavioral patterns and risk profiles change over time with additional longitudinal data. Recent work in AI, particularly continual and lifelong learning, demonstrates how adaptive model-updating strategies can address concept drift in evolving, safety-critical systems, offering a promising direction for future dementia risk prediction research [72].

These investigations were beyond the scope of the current study, which aimed to establish a transparent and well-calibrated baseline framework for dementia risk prediction using population-based survey data [73–75].

8. Conclusions

This study demonstrates that AI-enabled, interpretable ML models can forecast dementia risk using non-clinical, population-based data from NHATS. By combining behavioral, social, and technological variables with demographic and health indicators, the proposed framework provides a transparent, reproducible method for predicting dementia risk among community-dwelling older adults.

Among the evaluated algorithms, ensemble methods, especially XGBoost, showed the strongest overall performance by balancing discrimination, calibration, and robustness to class imbalance. Cross-validation confirmed that all models significantly outperformed a prevalence-based non-informative baseline, while nonparametric repeated-measures testing revealed statistically significant differences in calibration across models. Importantly, SHAP-based explainability analyses demonstrated stable and interpretable patterns of feature importance across algorithms, with technology use, outdoor activity, and social network size consistently identified as key predictors.

Instead of emphasizing clinical deployment or system-level impact, this study advances methodology by demonstrating how calibration-aware evaluation and cross-model explainability can be applied to longitudinal population health data. The results highlight the potential of interpretable ML to support transparent risk stratification and hypothesis generation in dementia research, while also stressing the importance of validating these models in independent cohorts and real-world settings.

Author Contributions: Conceptualization, A.A.; methodology, A.A. and M.G.R.; software, A.A.; validation, A.A. and M.G.R.; formal analysis, A.A.; investigation, A.A. and M.G.R.; Resources, A.A., M.G.R.; data curation, A.A. and M.G.R.; writing—original draft preparation, A.A.; writing—review and editing, A.A., M.G.R. and V.R.P.; Visualization, A.A.; supervision, V.R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study used publicly available, de-identified data from the National Health and Aging Trends Study (NHATS). Ethical approval was not required for this secondary analysis, as no identifiable human subject data were used.

Data Availability Statement: The data used in this study are available from the National Health and Aging Trends Study (NHATS) and can be accessed upon reasonable request, subject to the applicable data use agreements. No new datasets were generated for this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Full Term
AI	Artificial Intelligence
AUC	Area Under the Receiver Operating Characteristic Curve
Brier	Brier Score
CV	Cross-Validation
F1	F1 Score
ICT	Information and Communication Technology
KNN	K-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
NHATS	National Health and Aging Trends Study
PR	Precision–Recall
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
XGBoost	eXtreme Gradient Boosting

References

1. Dementia. Available online: <https://www.who.int/news-room/fact-sheets/detail/dementia> (accessed on 17 May 2025).
2. Nichols, E.; Steinmetz, J.D.; Vollset, S.E.; Fukutaki, K.; Chalek, J.; Abd-Allah, F.; Abdoli, A.; Abualhasan, A.; Abu-Gharbieh, E.; Akram, T.T.; et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019. *Lancet Public Health* **2022**, *7*, e105–e125.
3. Barrett, J.P.; Olivari, B.S.; Price, A.B.; Taylor, C.A. Cognitive decline and dementia risk reduction: Promoting healthy lifestyles and blood pressure control. *Am. J. Prev. Med.* **2021**, *61*, e157–e160.
4. Li, J.; Pandian, V.; Davidson, P.M.; Song, Y.; Chen, N.; Fong, D.Y.T. Burden and attributable risk factors of non-communicable diseases and subtypes in 204 countries and territories, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. *Int. J. Surg.* **2025**, *111*, 2385–2397.
5. Brain, J.; Kafadar, A.H.; Errington, L.; Kirkley, R.; Tang, E.Y.; Akyea, R.K.; Bains, M.; Brayne, C.; Figueredo, G.; Greene, L.; et al. What’s new in dementia risk prediction modelling? An updated systematic review. *Dement. Geriatr. Cogn. Disord. Extra* **2024**, *14*, 49–74.
6. Walters, K.; Hardoon, S.; Petersen, I.; Iliffe, S.; Omar, R.Z.; Nazareth, I.; Rait, G. Predicting dementia risk in primary care: Development and validation of the dementia risk score using routinely collected data. *BMC Med.* **2016**, *14*, 6.
7. Vermeulen, R.J.; Andersson, V.; Banken, J.; Hannink, G.; Govers, T.M.; Rovers, M.M.; Rikkert, M.G. Limited generalizability and high risk of bias in multivariable models predicting conversion risk from mild cognitive impairment to dementia: A systematic review. *Alzheimer’s Dement.* **2025**, *21*, e70069.
8. Rudin, C. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215.
9. Obermeyer, Z.; Emanuel, E.J. Predicting the future—Big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **2016**, *375*, 1216–1219.

10. Veronese, N.; Bolzetta, F.; Gallo, L.; Durante, G.; Vernuccio, L.; Saccaro, C.; Gambino, C.M.; Custodero, C.; Portincasa, P.; Morrotti, A.; et al. Clinical prediction models using artificial intelligence approaches in dementia. *Aging Clin. Exp. Res.* **2025**, *37*, 233.
11. Martin, S.A.; Townend, F.J.; Barkhof, F.; Cole, J.H. Interpretable machine learning for dementia: A systematic review. *Alzheimer's Dement.* **2023**, *19*, 2135–2149.
12. Piolatto, M.; Bianchi, F.; Rota, M.; Marengoni, A.; Akbaritabar, A.; Squazzoni, F. The effect of social relationships and cognitive decline in older adults: A systematic review and meta-analysis. *BMC Public Health* **2022**, *22*, 1–14. <https://doi.org/10.1186/s12889-022-12567-5>
13. Floud, S.; Balkwill, A.; Sweetland, S.; Brown, A.; Reus, E.M.; Hofman, A.; Blacker, D.; Kivimaki, M.; Green, J.; Peto, R.; et al. Cognitive and social activities and long-term dementia risk. *Lancet Public Health* **2021**, *6*, e116–e123.
14. Stern, Y. Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol.* **2012**, *11*, 1006–1012.
15. Benge, J.F.; Scullin, M.K. Technology use and cognitive aging: A meta-analysis. *Nat. Hum. Behav.* **2025**, *9*, 1405–1419.
16. Deng, C.; Shen, N.; Li, G.; Zhang, K.; Yang, S. Digital isolation and dementia risk in older adults: A longitudinal cohort study. *J. Med. Internet Res.* **2025**, *27*, e65379.
17. Lundberg, S.M.; Allen, P.G.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Available online: <https://github.com/slundberg/shap> (accessed on 17 May 2025).
18. Kavitha, C.; Mani, V.; Srividhya, S.R.; Khalaf, O.I.; Tavera Romero, C.A. Early-stage Alzheimer's disease prediction using machine learning. *Front. Public Health* **2022**, *10*, 853294.
19. Hossain, M.K.; Ashraf, A.; Islam, M.M.; Sourav, S.H.; Shimul, M.M.H. Optimizing Alzheimer's disease prediction through ensemble learning and feature interpretability with SHAP-based feature analysis. *Alzheimer's Dement. Diagn. Assess. Dis. Monit.* **2025**, *17*, e70162.
20. Fernández-Blázquez, M.A.; Ruiz-Sánchez de León, J.M.; Sanz-Blasco, R.; Verche, E.; Ávila-Villanueva, M.; Gil-Moreno, M.J.; Montenegro-Peña, M.; Terrón, C.; Fernández-García, C.; Gómez-Ramírez, J. XGBoost models based on non-imaging features for the prediction of mild cognitive impairment in older adults. *Sci. Rep.* **2025**, *15*, 29732.
21. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M.D.; et al. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.* **2022**, *23*, 1–61.
22. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832.
23. Ahmmad, M.R.; Hossain, E.; Khan, M.T.F.; Paudel, S. Using Machine learning to identify risk factors for Alzheimer's disease among older adults in the United States. *J. Alzheimer's Dis. Rep.* **2025**, *9*, 25424823251377691.
24. Gu, Z.; Liu, S.; Ma, H.; Long, Y.; Jiao, X.; Gao, X.; Du, B.; Bi, X.; Shi, X. Estimation of Machine learning-based models to predict dementia risk in patients with atherosclerotic cardiovascular disease: UK Biobank study. *JMIR Aging* **2025**, *8*, e64148.
25. Li, W.; Zeng, L.; Yuan, S.; Shang, Y.; Zhuang, W.; Chen, Z.; Lyu, J. Machine learning for the prediction of cognitive impairment in older adults. *Front. Neurosci.* **2023**, *17*, 1158141.
26. Benge, J.; Scullin, M.K. The technological reserve hypothesis: A meta-analysis. *Alzheimer's Dement.* **2023**, *19*, S23.
27. Hsu, E.-C.; Spaulding, E.M.; Jutkowitz, E. Technology activities and cognitive trajectories among community-dwelling older adults: NHATS. *JMIR Aging* **2025**, *8*, e77227.
28. Ren, X.; Qin, Y.; Li, B.; Wang, B.; Yi, X.; Jia, L. A core-space gradient projection-based continual learning framework under variable operating conditions. *Reliab. Eng. Syst. Saf.* **2024**, *252*, 110428.
29. Jeon, S.; Charles, S.T. Internet-based social activities and cognitive functioning two years later among middle-aged and older adults. *JMIR Aging* **2024**, *7*, e63907.
30. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56.
31. Yi, F.; Yang, H.; Chen, D.; Qin, Y.; Han, H.; Cui, J.; Bai, W.; Ma, Y.; Zhang, R.; Yu, H. XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 137.
32. Alatrany, A.S.; Khan, W.; Hussain, A.; Kolivand, H.; Al-Jumeily, D. An explainable machine learning approach for Alzheimer's disease classification. *Sci. Rep.* **2024**, *14*, 2637.
33. Kasper, J.D.; Freedman, V.A. Findings from the First Round of the National Health and Aging Trends Study (NHATS): Introduction to a Special Issue. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **2014**, *69*, S1–S7.
34. Freedman, V.A. Late-Life Disability and Care: An Update from the National Health and Aging Trends Study at Its 10-Year Mark. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **2022**, *77*, S1–S8.
35. Majlatow, M.; Shakil, F.A.; Emrich, A.; Mehdiyev, N. Uncertainty-Aware Predictive Process Monitoring in Healthcare: Explainable Insights into Probability Calibration for Conformal Prediction. *Appl. Sci.* **2025**, *15*, 7925.

36. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A Tutorial on Calibration Measurements and Calibration Models for Clinical Prediction Models. *J. Am. Med. Inform. Assoc.* **2021**, *27*, 621–633.
37. Yoon, H.-K.; Kim, B.R.; Kim, H.Y.; Park, D.K.; Kim, H.S.; Cho, H.-Y.; Lee, H.-C.; Lee, H. Multicenter Validation of a Scalable, Interpretable, Multitask Prediction Model for Multiple Clinical Outcomes. *npj Digit. Med.* **2025**, *8*, 583.
38. Baumgart, M.; Snyder, H.M.; Carrillo, M.C.; Fazio, S.; Kim, H.; Johns, H. Summary of the Evidence on Modifiable Risk Factors for Cognitive Decline and Dementia: A Population-Based Perspective. *Alzheimer's Dement.* **2015**, *11*, 718–726.
39. Yang, M.; Pajewski, N.; Espeland, M.; Easterling, D.; Williamson, J.D. Modifiable Risk Factors for Homebound Progression among Those with and without Dementia in a Longitudinal Survey of Community-Dwelling Older Adults. *BMC Geriatr.* **2021**, *21*, 561.
40. Ranson, J.M.; Rittman, T.; Hayat, S.; Brayne, C.; Jessen, F.; Blennow, K.; van Duijn, C.; Barkhof, F.; Tang, E.; Mummery, C.J.; et al. Modifiable Risk Factors for Dementia and Dementia Risk Profiling: A User Manual for Brain Health Services—Part 2 of 6. *Alzheimer's Res. Ther.* **2021**, *13*, 169.
41. Lin, L.; Wang, Y. SHAP Stability in Credit Risk Management: A Case Study in Credit Card Default Model. *Risks* **2025**, *13*, 238.
42. Givisis, I.; Kalatzis, D.; Christakis, C.; Kiouvrekis, Y. Comparing Explainable AI Models: SHAP, LIME, and Their Role in Electric Field Strength Prediction over Urban Areas. *Electronics* **2025**, *14*, 4766.
43. Sun, D.Q.; Huang, J.; Varadhan, R.; Agrawal, Y. Race and Fall Risk: Data from the National Health and Aging Trends Study (NHATS). *Age Ageing* **2016**, *45*, 120–127.
44. Long, E.; Gould, E.; Maslow, K.; Lepore, M.; Bercaw, L.; Leopold, J.; Lyda-McDonald, B.; Ignaczak, M.; Yuen, P.; Wiener, J.M.; et al. *Identifying and Meeting the Needs of Individuals with Dementia Who Live Alone*; U.S. Dep. Health Hum. Serv.: Washington, DC, USA, 2015.
45. Nakahara, K.; Yokoi, K. Role of Meaningful social participation and technology use in mitigating loneliness and cognitive decline among older adults. *Am. J. Occup. Ther.* **2024**, *78*, 7806205150.
46. Liu, Y.; Lu, G.; Liu, L.; He, Y.; Gong, W. Cognitive reserve over the life course and risk of dementia: A systematic review and meta-analysis. *Front. Aging Neurosci.* **2024**, *16*, 1358992.
47. Meng, X.; D'Arcy, C. Education and dementia in the context of the cognitive reserve hypothesis: A systematic review and meta-analysis. *PLoS ONE* **2012**, *7*, e38268.
48. Fancourt, D.; Steptoe, A.; Cadar, D. Community engagement and dementia risk: Time-to-event analyses from a national cohort study. *J. Epidemiol. Community Health* **2020**, *74*, 71–77.
49. Chen, Y.; Grodstein, F.; Capuano, A.W.; Wang, T.; Bennett, D.A.; James, B.D. Late-life social activity and subsequent risk of dementia and mild cognitive impairment. *Alzheimer's Dement.* **2025**, *21*, e14316.
50. James, B.D.; Wilson, R.S.; Barnes, L.L.; Bennett, D.A. Late-life social activity and cognitive decline in old age. *J. Int. Neuropsychol. Soc.* **2011**, *17*, 998–1005.
51. Crooks, V.C.; Lubben, J.; Petitti, D.B.; Little, D.; Chiu, V. Social network, cognitive function, and dementia incidence among elderly women. *Am. J. Public Health* **2008**, *98*, 1221–1227.
52. Rocha, R.; Fernandes, S.M.; Santos, I.M. Technology-assisted cognitive and physical interventions in older adults: A systematic review. *Healthcare* **2023**, *11*, 2375.
53. Zhang, J.; Wang, J.; Liu, H.; Wu, C. Association of dementia comorbidities with caregivers' physical, psychological, social, and financial burden. *BMC Geriatr.* **2023**, *23*, 60.
54. Sacchetti, S.; Locatelli, G.; Altomare, D.; Guaita, A.; Rolandi, E. Evidence on protective factors for dementia and cognitive impairment in older adults: An umbrella review. *Dement. Geriatr. Cogn. Disord.* **2025**, *54*, 394–408.
55. Chen, B.; Yang, C.; Ren, S.; Li, P.; Zhao, J. Internet use and cognitive function among middle-aged and older adults. *J. Med. Internet Res.* **2024**, *26*, e57301.
56. Taniguchi, R.; Ukawa, S. Participation in social group activities and risk of dementia. *Open Public Health J.* **2022**, *15*, 1–10.
57. Cho, E.; Kim, S.; Heo, S.J.; Shin, J.; Hwang, S.; Kwon, E.; Lee, S.; Kim, S.; Kang, B. Machine learning-based predictive models for the occurrence of behavioral and psychological symptoms of dementia. *Sci. Rep.* **2023**, *13*, 8073.
58. Abegaz, T.M.; Ahmed, M.; Ali, A.A.; Bhagavathula, A.S. Predicting health-related quality of life using social determinants of health: A machine learning approach with the all of us cohort. *Bioengineering* **2025**, *12*, 166.
59. Olson, R.S.; La Cava, W.; Mustahsan, Z.; Varik, A.; Moore, J.H. Data-driven advice for applying machine learning to bioinformatics problems. *Pac. Symp. Biocomput.* **2018**, *23*, 192–203.
60. Yuan, L.; Zhang, Y.; Wu, Y.; Zhang, A.; Bai, H.; He, M.; Wang, Z.; Zheng, L. Machine learning-based early screening of mild cognitive impairment using nutrition-related biomarkers and functional indicators. *Front. Aging Neurosci.* **2025**, *17*, 1641690.

61. Akter, S.; Guess, T.M.; Sarker, S.; Hocket, S.A.; Kiselica, A.M.; Hall, J.B.; Rao, P. Explainable machine learning for early detection of mild cognitive impairment, fall risk, and frailty using sensor-based motor function data. *medRxiv* **2025**, <https://doi.org/10.64898/2025.12.23.25342943>.
62. He, Y.; Leng, Y.; Vranceanu, A.M.; Ritchie, C.S.; Blacker, D.; Das, S. Predictive model for cognitive decline using social determinants of health. *J. Aging Res. Lifestyle* **2026**, *15*, 100056.
63. Hasanzadeh Khosroshahi, M.; Morasi, S.; Gharkhanlou, S.; Matamedi, A.; Hassannabghlou, S.; Vahedi, H.; Pedrammehr, S.; Kabir, H.M.D.; Jafarizadeh, A. Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. *Diagnostics* **2025**, *15*, 612.
64. Lee, J.Y.; Lee, S.Y. AI-based predictive algorithm for early diagnosis of high-risk dementia groups. *Healthcare* **2024**, *12*, 1872.
65. Arnaud, É.; Moreno-Sanchez, P.A.; Elbattah, M.; Ammirati, C.; van Gils, M.; Dequen, G.; Ghazali, D.A. Development and clinical interpretation of an explainable AI model for predicting patient pathways in the emergency department: A retrospective study. *Appl. Sci.* **2025**, *15*, 8449.
66. Netayawijit, P.; Chansanam, W.; Sorn-In, K. Interpretable machine learning framework for diabetes prediction. *Healthcare* **2025**, *13*, 2588.
67. Cui, X.; Zheng, X.; Lu, Y. Prediction model for cognitive impairment among disabled older adults. *Healthcare* **2024**, *12*, 1028.
68. Pamias-Lopez, D.; Keck, T. Healthy ageing centres in Bosnia and Herzegovina is associated with increased physical activity, social interactions, and life satisfaction among older people. *J. Ageing Longev.* **2025**, *5*, 5.
69. Zhang, M.; Cui, Q.; Lü, Y.; Yu, W.; Li, W. Multimodal learning framework for Alzheimer's disease diagnosis. *Comput. Ind. Eng.* **2024**, *197*, 107968.
70. Henríquez, P.A.; Araya, N. Multimodal Alzheimer's disease classification via ensemble neural networks. *PeerJ Comput. Sci.* **2024**, *10*, e2590.
71. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232.
72. Besnard, Q.; Ragot, N. Continual Learning for Time Series Forecasting: A First Survey. *Eng. Proc.* **2024**, *68*, 49.
73. Alsubaie, M.G.; Luo, S.; Shaukat, K. Alzheimer's disease detection using deep learning on neuroimaging: A systematic review. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 464–505.
74. Zhou, Q.; Wang, J.; Yu, X.; Wang, S.; Zhang, Y. Deep learning for mathematical reasoning. *Proc. Annu. Meet. Assoc. Comput. Linguist.* **2023**, *1*, 14605–14631.
75. De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; Tuytelaars, T. A Continual learning survey: Defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3366–3385.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.