AGENTIC AI, MACHINE LEARNING

Explainable Artificial Intelligence for Longitudinal Prediction of Cognitive Decline in Diabetes: A Comprehensive Survey of Methods, Data, and Research Challenges

Jitendra Kumar Sonkar[1], Vishal Kohli[1], Sachin Kumar[2]

[1]Department of Computer Science and Engineering, Neelkanth Institute of Technology, Meerut, Uttar Pradesh, India
[2]Department of Computer Science and Engineering, IEC College of Engineering & Technology, Greater Noida, Uttar Pradesh, India

Published November 2026

Abstract

Although diabetes mellitus is linked to an increased risk of cognitive impairment and dementia, individual courses of these disorders vary and are not easily foretold. The present survey investigates epidemiological evidence, a longitudinal cohort design, machine learning (ML), explainable artificial intelligence (XAI), and reliability needs of prediction system for cognitive decline in individuals with diabetes. The Health and Retirement Study (HRS) is highlighted, as are repeated cognitive assessments, diabetes diagnosis, limited biomarker information, attrition, and the difference between prevalent impairment and incident decline. We will compare statistical longitudinal models, tree ensembles, survival approaches, temporal neural networks, feature attribution methods and calibrated risk estimation. The identified methodological issues are outcome inconsistency, information leakage, small biomarker linked subcohorts, confounding, selective follow-up, unstable explanations, and weak transportability, and are typical features of a structured evidence synthesis. We describe then a research framework based on evaluation, which integrates participant level and temporal validation, multi-horizon outcomes, explanation stability, calibration, and assessment of subgroups. This article is a narrative methodological survey and not a meta-analysis; there is no training of a new clinical model and no performance claims are made. The review offers a practical base for repeatable and judicious studies for explainable prediction of diabetes associated cognitive decline.

Keywords: Diabetes mellitus; cognitive decline; dementia; Health and Retirement Study; longitudinal machine learning; explainable artificial intelligence; SHAP; calibration; temporal validation

1 Introduction

Diabetes mellitus (DM) has been shown to have a negative effect in multiple organ systems and has been linked in longitudinal observational studies to differences in cognitive performance and subsequent cognitive decline [1]–[4]. Cognitive ageing, on the other hand, is not a disease process in itself. Memory, attention, processing speed and executive function can all develop differently, and the instrument used to measure, the timeframe of the assessment, and the prior level of ability all have an impact on what is considered a decline. Therefore, predictive modelling should start with a precise definition of the target outcome which should not be synonymous with dementia, mild cognitive impairment and changes in scores at the time of the short term evaluation.

Age, education, vascular disease, depression, glycaemic measures and prior cognition are potentially informative variables identified in both clinical and population-based studies. Not all change in a

predictor will lead to change in cognitive outcome if there is an association. While machine learning can capture interactions and nonlinearities, it cannot, on its own, remove the confounding or selection bias. An apparently accurate black-box model may be ineffective in high stakes environments if the probabilities are wrong, the model provides explanations that change over time, or if the validation relies on information from subsequent visits.

In particular, the Health and Retirement Study (HRS) is particularly relevant in that it tracks older adults over time via repeated interviews and cognitive assessments [5], [6]. It also represents practical issues: of intermittent participation, proxy answers, mortality, and different availability of measurement outcomes - either a subset or a full cohort of patients - with laboratory biomarkers. Predictions for diabetes must thus be developed in a way that relies on clear specifications of the cohort, careful selection of outcomes and consideration of missingness, and is made transparent.

The overall structure of this survey is similar to the comprehensive IEEE-style survey: Background, Literature Review, Analytical Approaches, Practical Support and Decisions use considerations, Comparative Assessment, Research gaps and future directions. It also covers prediction design for longitudinal prediction, harmonisation of HRS variables, fidelity of interpretability, fairness, calibration, and responsible deployment. The primary focus in this case is on methodological evaluation of predictions systems, not medical treatment.

2 Overview of diabetes-associated cognitive decline

2.1 Clinical and epidemiological context

A number of factors, such as ageing, vascular risk and socioeconomic factors, are shared by diabetes and cognitive impairment. Long-term follow-up prospective studies have reported associations between diabetes and cognition [1], [2]. HbA1C and cognition studies suggest that the exposure can be different among participants and over time, and that the results are influenced by treatment, comorbidities, age and selection for biomarker sampling [3], [4].

Cognitive decline is a change within an individual, while cognitive impairment is a cognitive state, with a score threshold or clinical classification. To make a diagnosis of dementia, more information about functional and clinical abilities is needed than a single cognitive score. Three different tasks, but related: Continuous change for machine learning, reliable threshold for decline, incident impairment, and time to diagnosis. Comparisons of models are only useful if the definitions of the outcomes and follow-up windows are comparable.

2.2 Cognitive domains and measurement

HRS assess measures cognition with interview-based tasks, including word recall, and other mental-status tasks, using different instruments and harmonized measures across waves and eligibility for each [5], [6]. Immediate and delayed recall mainly assess episodic memory, orientation, serial subtraction and related items assess other cognitive functions. It is helpful to include total scores, but it can mask domain-specific change. Before a binary decline label is built, ceiling effects, practice effects, sensory limitations, and proxy interviews and measurement non-equivalence across waves have to be taken into account.

2.3 Risk factors or confounding

Possible predictors include age, level of education, baseline cognition, blood pressure, prior stroke, cardiovascular disease, depressive symptoms, physical functioning and smoking and activity patterns. Diabetes markers can include self-reporting of a physician diagnosis, treatment status and selected-wave HbA1c. Some factors are both predictor and consequence of cognitive decline and it is important for the timing of them. The availability of the predictors at the decision date will need to be documented. High model importance should not be interpreted as causal contributions by associations.

3 Literature review

3.1 Longitudinal diabetes–cognition evidence

Evidence of diabetes–cognition effects is available in a longitudinal form.

The work by Marden and colleagues was directly relevant to the proposed cohort and exposure definition, as they used data from the HRS to examine HbA1c, diabetes and memory decline [3]. The study of Rawlings and colleagues provided an example of the need for long follow-up and early life metabolic exposure for diabetes in midlife and cognitive change over time [2]. Zheng and coworkers analyzed HbA1c, diabetes and cognitive decline in the English Longitudinal Study of Ageing, which was a separate but relevant ageing population-based study [4]. These studies stimulate risk assessment but are not a validated individual-level forecasting tool.

Biological factors include vascular injury, insulin signalling, inflammation and glycaemic dysregulation. However, it is important to distinguish between mechanistic plausibility and observational association and between individual prognosis. Whether the variable is associated with a better discrimination and calibration than other variables available to the predictive model, such as age, baseline cognition, is the question, not whether SHAP provides a biological mechanism.

3.2 Machine learning for cognitive prediction

The authors of the current paper, Aschwanden et al., have modeled multiple risk factors together in a longitudinal design to predict cognitive impairment and dementia in an ageing population through the use of machine learning and survival analysis [7]. This paper provides a methodological reference for time to event endpoints and comparative baseline models. It is not however interchangeable with a diabetes restricted cohort, since the broader older-adult population is not a comparable outcome to a change-score outcome, and a different validation protocol from incident dementia is required.

Baseline models for repeated cognitive scores are interpretable models for linear and mixed-effects regression. When time-to-event outcomes are credible and event and censoring definitions are credible, Cox proportional-hazards models and survival forests work well. Non-linearities in tabular predictors can be captured by random forests, gradient boosting and regularized generalized linear models. Recurrent networks and transformers can be used to model sequences, but can overfit sparse and irregular waves, and should not be assumed to be superior to well-tuned tabular models. To allow for comparison, all families must be matched in train/ test cohorts, as well as having the same timeframe.

3.3 Explainable AI foundations

Lundberg and Lee proposed SHAP as a general method for explaining predictions by features [8]. LIME for local, model-agnostic approximations was proposed by Ribeiro et al. [9]. They are both useful for auditing the behaviour of the model, but both rely on reference distributions and the choice of features and perturbations. High attribution does not necessarily mean it represents an actionable risk factor; it could be a proxy for age or survey wave or for prior cognition.

Explanations should be evaluated for stability, consistent and clinical plausibility. There are cases where correlated features can share the importance; missing-value indicators can pull the weight; explanations can change once they've been recalibrated or retrained. Local explanations should not be offered as "how to" information. Global and local attribution, uncertainty, explanation stability under bootstrapping and sensitivity to the removal of the feature groups are useful reportings.

4 Data sources and longitudinal analysis

4.1 Health and Retirement Study (HRS)

HRS is a nationally oriented US longitudinal study of ageing, health, retirement and family circumstances. RAND harmonized longitudinal products facilitate cross-wave analysis, while cognition documentation clarifies test construction and imputed measures [5], [6]. Researchers should consult the specific release codebook and usage terms before choosing variable names. Public-use self-reported diabetes is not a guaranteed type-2 diabetes diagnosis; subtype claims require separate evidence.

The core analytic table should have one row per eligible participant per prediction origin. At baseline t, include only observations recorded by t; construct outcomes from cognitive measurements at t+2 and t+4 years, subject to wave availability and prespecified tolerance. Multiple origins from one person must never be divided across training and test partitions. When survey design weights are used, their relevance to the analytic subcohort and prediction objective should be justified.

4.2 Biomarker and missing-data limitations

Not all laboratory biomarkers are measured at all HRS waves. A complete-case biomarker analysis could yield a much smaller and biased sample than a model based on survey variables. It is therefore preferable to have a two-tier design with a broad core model based on diabetes status, cognition and routine covariates, and a biomarker-augmented model on the eligible linked subset. Absolutely not, for differences in performance between these tiers cannot be explained from an HbA1c level, unless both tiers are measured in the same participants.

4.3 Outcomes, censoring and attrition

Follow-up cognitive score – baseline score may be a continuous outcome, accounting for harmonization and practice effects. There must be a defensible threshold in a binary endpoint, ideally based on published reliable change criterion or a prespecified criterion for training only. Death and loss to follow-up are not necessarily random. Participant flow, number of events, missingness across the waves, and sensitivity analysis (inverse-probability weighting or competing-risk if appropriate) should be reported.

4.4 Leakage-safe validation

The method needs to learn only when the training folds; and the parameters of imputation, standardization, feature selection, class weighting and hyperparameter tuning. Participant-disjoint cross validation is for generalisation to unseen people while a later-wave holdout tests for temporal robustness. The test set should not be used for guiding the selection of thresholds, calibration fitting or model selection. When the data is repeated, it needs to be split, and when the outcome windows overlap, there must be explicit splitting rules.

5 Predictive and explainability approaches

5.1 Statistical and ML baselines

A defensible baseline ladder starts with age, sex, education, and baseline cognition and then vascular and lifestyle features, and then diabetes-specific features. A simple benchmarking method is linear or logistic regression, and random forest and gradient boosting tests a nonlinear signal while repeated measures are accounted for with mixed-effects models. If dates and censoring are well-determined, then survival analysis is appropriate for incident impairment. It is important that hyperparameter budgets are similar between methods and preprocessing should be similar.

5.2 Multi-horizon prediction

Separate 2-year and 4-year models are easy to assess; a common model can leverage across follow-up windows by conditioning on the horizon. Reporting should differentiate between a continuous score change and binary decline risk assessment. Classification can be measured by ROC-AUC, PR-AUC, sensitivity, specificity and F1, operating-point behaviour can be measured by sensitivity, specificity and F1, and probability quality can be measured by Brier score, calibration intercept/slope and reliability plots. Uncertainty intervals should be used in addition to MAE and RMSE to convey the concept of a continuous change.

5.3 Explanation methods and evaluation

They are able to explain how to evaluate. They know how to evaluate. The predictors that drive a trained model can be identified via SHAP summary plots and dependence analyses, which can be complemented by a local approximation obtained using LIME. The potential of feature interactions across age and baseline cognition and glycaemic indicators could be explored, but tested for robustness. Stability of explanation can be assessed based on agreement of top-k ranked features in repeated bootstrap retraining; variability of attributions for matched records; or subgroup-specific rank concordance. Stability is not synonymous with correct or causal.

5.4 Reliability, fairness and uncertainty

Validation data should be used to estimate calibration, and a separate untouched test data set used to check the calibration. Subgroup analyses by age, sex and education may show different error rates or poor calibration, in which case they should show uncertainty intervals instead of definite claims of fairness. Performance differences may occur due to the absence of data or due to the mode of the survey and attrition. An uncertainty-aware system can indicate uncertainty and out-of-distribution or incomplete records and avoid giving unsupported high confidence assessments.

6 Support approaches and responsible decision use

Unlike an autism survey on behavioural issues, the domain of autism support addressed in this study relates to research-based cognitive risk communication. A prediction tool might help in the screening research at the population level or help prioritise those who need further screening but should not be used independently to diagnose dementia or to suggest glucose treatment in place of clinical judgement. Risk estimates should mention the time horizon of the prediction, the definition of the outcome, the reference population and risk uncertainty/limitation. A usable interface might show the expected probability of a prespecified decline event along with the assumptions made in the baseline or follow-up periods, as well as the most important explanatory factors, and a reminder that associations are not causal. The display needs to also show if they do not match the training population and whether there is an evaluation of the model calibration for their subgroup. A clinician facing explanation vs a patient facing explanation: the latter must be worded carefully to prevent unnecessary alarm. Sensitive data like ageing, diabetes and cognitive data imply privacy and governance is an integral part. Research workflows should meet the access requirements of the HRS, and should not include a method for publishing record-level identifiers and document reproducibility without revealing participants. A transparent model card should provide an explanation on population, exclusions, measurement windows, metrics, subgroup limitations, intended uses and unintended uses. Validation, impact and prospective evaluation, outside of a retrospective thesis, would be needed to achieve clinical adoption.

7 Comparative analysis and research gaps

7.1 Cross-study comparison

The literature provides answers to questions, not answers to single questions that are interchangeable. Epidemiological studies estimate associations and cognitive trajectories [2], [3]; survival ML studies forecast incident impairment or dementia [7]; XAI methods give post hoc model descriptions [8], [9]. They have different endpoints, cohorts, follow-up periods and metrics. Without a standardized experimental protocol, there is no basis for claiming one algorithm is superior to another, across papers. Table 1 summarises the evidence streams.

Table 1. Cross-study comparison.

Evidence stream	Strength	Limitation for thesis
HRS diabetes-memory [3]	Relevant exposure and cohort	Association, not full XAI forecast
Long-term cohorts [2],[4]	Repeated cognition and glycaemia	Different populations/targets
HRS ML survival [7]	Longitudinal prediction methods	Not diabetes-specific
SHAP/LIME [8],[9]	Local/global interpretation	Attribution not causation
TRIPOD [10]	Transparent reporting guidance	Not proof of clinical utility
PROBAST [11]	Bias/applicability assessment	Requires study-specific judgments

7.2 Research gaps

- **Gap 1: Heterogeneity of outcome and heterogeneity of horizon.** A variety of cognitive instruments or incident-dementia definitions are frequently used in concert in studies. For future

studies, it is important to evaluate the clearly specified 2- and 4-year outcomes separately and to see both outcomes in the same context, where eligibility requirements are the same.

- **Gap 2: Diabetes-specific generalization.** Ageing models in general may not be applicable without modification to a diabetes-specific population. It should report the size of the diabetic cohort, the prevalence of diabetes in the cohort, and the outcomes of the covariate shifts and the performance of various subgroups, but should not claim novelty based on limiting a dataset.
- **Gap 3: Biomarker-linked selection.** Selected participants and waves are observed for HbA1c. The incremental value should be assessed in the common linked subset of the study having a matched non-biomarker baseline, and the missingness and selection should be described.
- **Gap 4: Temporal leakage and attrition.** Performance can be exaggerated due to overlapping records, imputation from future waves and selective survival. Disjoint holdout and later-wave holdout validation, accounting for attrition and sensitivity analysis are required.
- **Gap 5: Explanation instability.** The important thing with feature importance is to have it reported without considering the variation of the feature over the samples, waves and subgroups. This uncertainty could be quantified by using Bootstrap ranking agreement and attribution sensitivity approaches.
- **Gap 6: Calibration and usefulness for decision making.** Trustworthy individual risk is not a product of discrimination. Before even considering decision support, Brier score, calibration plots, confidence intervals and abstention analyses are required.
- **Gap 7: External validity and reporting.** HRS is a specific context of ageing. Limitation can be made visible through transparent reporting, as outlined in the TRIPOD and risk-of-bias principles in PROBAST principles; true external validation through independent cohort.

8 Proposed future research framework

8.1 Cohort-to-model pipeline

A proposed reproducible workflow is: Select eligible HRS participants with baseline diabetes; Harmonize cognitive tests; Define 2- and 4-year outcomes; Develop baseline-only predictors; Partition and hold out rest of the HRS waves; Preprocess within training; Compare statistical and ML models; Calibrate predictions; Audit explanations; Quantify sensitivity of subgroups and attrition; Document results. The proposal is not a real or tested algorithm, but an assessment plan.

8.2 Experimental design and ablation

Ablations should be compared between demographics, demographics and baseline cognition, demographics and clinical/lifestyle history, and matched subset with demographics with biomarker information. Make comparisons between SHAP of original and bootstrap trained models and report feature rank concordance and confidence intervals. If an explicit optimization method and independent benchmark demonstrate the value of a stability-aware regularization term, then that term can be considered.

8.3 Reproducibility and scientific boundaries

Pre-register outcome definitions, prediction origins, missing-data policies and primary metrics, if applicable. Provide release code, cohort-construction scripts and synthetic data instead of limited participant data. If the result is a statistical confidence interval, uncertainty in selection, and a

negative finding should be reported. The effects of interventions on glucose control cannot be designed as predictions of preventing decline.

9 Practical evaluation checklist and reporting

9.1 Dataset provenance and cohort flow

A research ready paper should specify the exact HRS release, the date of downloading, the age included, the waves in the survey, the diabetes definition, and the derivation of the cognitive score, as well as the permitted missingness. A flow diagram should depict the number of unique individuals before and after exclusions, number of individuals with cognition available at baseline, number at each of the follow-up time horizons, and number that could be linked to HbA1c. Unit of analysis has to be specified: person, person-wave or prediction origin. Effective sample size can be considerably smaller than the total number of longitudinal records, due to the fact that repeated observations are correlated.

It is important for the analyst to differentiate between survey nonresponse and meaningful missingness. Illness, institutionalization, death or transition to proxy reporting may also be reasons for missing cognitive interviews. Routine complete case exclusion may lead to a bias of the study population of higher risk. For sensitivity analysis, comparisons should be made between those who were retained and those who were excluded, with regard to measured baseline characteristics; the extent of follow-up should be reported in a table, and results should be examined under reasonable assumptions about missingness.

9.2 Feature availability and temporal provenance

There should be a provenance table for each feature containing the variable label, source product, wave, observation time, encoding, missingness percentage and justification. Diabetic status may alter with time and therefore a baseline diabetic cohort should not be based on the diabetes diagnosis after the prediction origin. The same applies to stroke, treatment variables and depression. The prediction date is no longer a row index but actual information.

If sufficient historical observations exist, feature engineering can include prior cognitive slopes, variability of historical scores, length of time since reported diagnosis of diabetes, and cumulative vascular burden. Please note that engineered variables must be based on data collected prior to baseline. Predictors should not include any indicators measuring the eventual outcome interview or future missingness, since they would serve as a proxy for the eventual outcome, rather than measuring cognition.

9.3 Statistical reporting and model comparisons

All models should have confidence intervals and evaluated on the same denominator. When interpreting AUC differences, they should be taken with a grain of salt and a small improvement could be statistically or clinically insignificant. Precision–recall and positive predictive value at prespecified sensitivity levels may be more informative than accuracy for imbalanced decline events. Choosing thresholds based on test data will introduce bias in the estimates of performance.

This is especially relevant since the point of risk prediction is not to merely rank the risks but to predict the probability of those risks. High AUC does not guarantee that a model is not systematically over-estimating the risk for a person to experience decline. Observed versus predicted frequencies

with uncertainty bands should be plotted in calibration plots and the Brier score compared with a simple reference based on the prevalence of the forecasted frequency. If a calibration model is being fitted, it is part of the training/validation set, and should not be part of the held out test set.

To understand the meaning of MAE and RMSE for continuous cognitive change, they need to be interpreted in relation to the score scale, measurement reliability, and a simple prediction of no change. For models that are not superior to predicting the baseline score or age-adjusted linear trend, the added value of complex machine learning is minimal. Ideally, baselines should be firm enough to ensure credibility of positive results.

9.4 Robustness, subgroup evaluation and explanations

Explanation stability needs to be assessed through repeated bootstrap training and small perturbations of noncritical inputs and also separate temporal validation sets. By using rank correlation, the top-k overlap and variability in local attribution magnitude, one can show when a feature-importance narrative is fragile. The chosen reference background for SHAP and how correlated variables are treated should be made public. Never read explanations as a "list of interventions" or "ranked list of intervention.

Wherever possible, results should be presented for subgroups, including age bands, sex, and education. An uncalibrated AUC may mask systematic overconfidence for certain participants. Differences should be viewed with caution due to the sources of the surveys, definitions of events, and measurement tools used. When dealing with small groups the uncertainty may be too broad to draw conclusions, and a clear expression of insufficient evidence is better than a false sense of precision.

9.5 Threats to validity

Leakage, inconsistent time windows, unrecorded treatment changes and proxy responses and misclassification of outcomes are threats to internal validity. Cohort geography, age distribution in cohort, cohort recruitment and follow-up may restrict the external validity. The construct validity of the cognitive measure selected relies on the ability to measure the intended concept of decline, not on the natural variability of tests. The validity of interpretation is compromised when feature attribution is used to support causation.

Survivor bias is an important problem. Participants who develop severe diabetes complications may die or drop out of the study or not participate in follow-up, altering the distribution of risk between those who had cognition observed. Mortality and dropout rates should be described in a study, and inverse-probability methods, joint models or competing-risk analysis should be considered. The methods rely on assumptions and do not ensure that bias is removed.

10 Discussion

The best chance for success is the M.Tech CSE thesis is not just the application of XGBoost and plotting SHAP values. The methodological contribution must be showcased in the context of leakage-aware temporal evaluation, matched feature ablations, explanation stability and calibrated multi-horizon prediction. These elements tie ML engineering to real-world constraints of longitudinal health data studies.

A survey also should acknowledge what has not yet been determined. Other papers reviewed vary in cognitive measures, methods of ascertainment for diabetes, windows of observation, and targets assessed. The present article does not quantitatively pool their effect sizes or search for databases or replicate results. It does not therefore claim that all the gaps indicated have not been found in all published work, but rather that the gaps indicated are the most plausible ones.

The use of external cohorts, causal inference designs, uncertainty-aware explanations and more complex temporal models could be explored in future investigations. It is okay to have complex ideas as long as there is an identifiable benefit, not trend appeal. A simpler "calibrated" model, with clear explanations and limited scope, might be better than an opaque model with peanuts increments in discrimination.

11 Conclusion

This survey examined the connection between diabetes-associated cognitive decline, longitudinal ageing data and machine learning and explainable AI. HRS provides the relevant basis for repeated cognitive assessments; however, availability of biomarkers and outcome harmonization, as well as time indexing and attrition, makes prediction at the individual level difficult. Research and literature are motivational for predictive modelling, interpretability tools are important to provide stability of explanation, and prediction-reporting standards emphasize the need to consider calibration and bias assessment. A future study that addresses those points should include multiple horizon baselines, both participant-disjoint and temporal, measure incremental feature value, and report uncertainty, but not make any claims of causality or clinical readiness. These requirements can be defended as a basis for an explainable and reliability-aware M.Tech research programme.

References

[1] G. J. Biessels and F. Despa, "Cognitive decline and dementia in diabetes mellitus: mechanisms and clinical implications," *Nature Reviews Endocrinology*, vol. 14, pp. 591–604, 2018. doi:10.1038/s41574-018-0048-7.

[2] A. M. Rawlings et al., "Diabetes in midlife and cognitive change over 20 years: a cohort study," *Annals of Internal Medicine*, vol. 161, pp. 785–793, 2014. doi:10.7326/M14-0737.

[3] J. R. Marden et al., "High hemoglobin A1c and diabetes predict memory decline in the Health and Retirement Study," *Alzheimer Disease & Associated Disorders*, 2017. doi:10.1097/WAD.0000000000000182.

[4] F. Zheng et al., "HbA1c, diabetes and cognitive decline: the English Longitudinal Study of Ageing," *Diabetologia*, vol. 61, pp. 839–848, 2018. doi:10.1007/s00125-017-4541-7.

[5] Health and Retirement Study, "Cognition data," *University of Michigan*. https://hrs.isr.umich.edu/data-products/cognition-data (accessed Nov. 2026).

[6] RAND, "RAND HRS Longitudinal File 2022," *HRS Data Products*. https://hrsdata.isr.umich.edu/data-products/rand-hrs-longitudinal-file-2022 (accessed Nov. 2026).

[7] D. Aschwanden et al., "Predicting cognitive impairment and dementia: A machine learning approach," *Journal of Alzheimer's Disease*, 2020. Full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC7934087/.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," *Proc*. ACM SIGKDD, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[10] G. S. Collins et al., "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," *Annals of Internal Medicine*, vol. 162, pp. 55–63, 2015. doi:10.7326/M14-0697.

[11] R. F. Wolff et al., "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Annals of Internal Medicine*, vol. 170, pp. 51–58, 2019. doi:10.7326/M18-1376.

[12] E. W. Steyerberg et al., "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, vol. 21, pp. 128–138, 2010. doi:10.1097/EDE.0b013e3181c30fb2.

[13] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, pp. 565–574, 2006. doi:10.1177/0272989X06295361.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. doi:10.1023/A:1010933404324.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc*. ACM SIGKDD, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[16] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, 2017.

[17] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society Series B*, vol. 34, pp. 187–220, 1972.

[18] H. Ishwaran et al., "Random survival forests," *Annals of Applied Statistics*, vol. 2, pp. 841–860, 2008. doi:10.1214/08-AOAS169.

[19] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge University Press, 2009.

[20] M. van Smeden et al., "Sample size for binary logistic prediction models: Beyond events per variable criteria," *Statistical Methods in Medical Research*, vol. 28, pp. 2455–2474, 2019. doi:10.1177/0962280218784726.